ZICO: A Credit Scoring Approach to Detecting Zombie Papers

Christophe Hurlin
Marc Joëts
Valérie Mignon

2026-18 Document de Travail/ Working Paper

EconomiX - UMR 7235 Bâtiment Maurice Allais
Université Paris Nanterre 200, Avenue de la République
92001 Nanterre Cedex

Site Web : economix.fr
Contact : secreteriat@economix.fr

ZICO: A Credit Scoring Approach to Detecting Zombie Papers*

Christophe Hurlin† Marc Joëts‡ Valérie Mignon§

July 17, 2026

Abstract

This paper proposes a new framework for assessing scientific credibility using probabilistic retraction risk. We introduce the Zombie Integrity and Credibility Oversight (ZICO) framework, a multidimensional credibility score inspired by credit-scoring models in finance. Rather than relying on isolated indicators of misconduct, ZICO integrates heterogeneous signals extracted from the semantic characteristics of scientific manuscripts and the broader organization of scientific production to estimate the *ex ante* probability that a publication will eventually become problematic or be retracted. We apply the framework to a corpus of publications in economics and finance. Our findings show that retraction risk is driven primarily by linguistic and semantic characteristics rather than by conventional metadata, such as authorship or institutional affiliations. In particular, measures of readability, lexical diversity, textual complexity, and citation practices emerge as the strongest predictors of scientific credibility. These signals capture deeper structural and stylistic irregularities rather than differences in English language proficiency, thereby highlighting semantic organization as a robust marker of research integrity. We further develop a dynamic version of ZICO that continuously updates credibility assessments, allowing the framework to operate as an adaptive early-warning system for editorial screening. More fundamentally, our framework shifts the assessment of scientific integrity from the *ex post* detection of misconduct to the *ex ante* probabilistic evaluation of manuscript credibility. Beyond its predictive performance, ZICO provides a transparent and interpretable framework for editorial decision-making, research evaluation, and the governance of research integrity.

JEL Classification: C38; C53; D83; A11; O33.

Keywords: Zombie papers; Research integrity; Retraction risk; Scientific publishing; Natural language processing; Machine learning.

* *Corresponding author*: Valérie Mignon, EconomiX-CNRS, University of Paris Nanterre, 200 avenue de la République, 92001 Nanterre cedex, France. E-mail: valerie.mignon@parisnanterre.fr

† University of Orléans (LEO) and Institut Universitaire de France. E-mail: christophe.hurlin@univ-orleans.fr

‡ IESEG School of Management; Univ. Lille, CNRS UMR 9221 - LEM Lille Economie Management F-59000 Lille, France. E-mail: m.joets@ieseg.fr

§ EconomiX-CNRS, University of Paris Nanterre and CEPII, Paris, France. E-mail: valerie.mignon@parisnanterre.fr

1 Introduction

While science fundamentally relies on the integrity of its knowledge base, the scientific publishing system has faced an increasing crisis of credibility and governance in recent years (Ioannidis et al. (2025)). Retractions have risen substantially across disciplines, and several prominent scholars have been exposed for scientific misconduct, including data and results fabrication, image manipulation, paper mills, plagiarism, and peer-review manipulation as well as for major methodological or analytical errors (Grieneisen & Zhang (2012), Steen et al. (2016), Hesselmann et al. (2017), Oransky (2022), Van Noorden (2023), Ioannidis et al. (2025)). High-profile cases such as social psychologist Diederik Stapel, whose fabricated datasets affected dozens of publications, or Francesca Gino in management research, whose work became the subject of extensive allegations of data manipulation, have reinforced concerns regarding the effectiveness of existing editorial and peer-review safeguards. Similar controversies have also emerged in economics and finance, including the Reinhart and Rogoff austerity debate, where analytical and coding errors in highly influential empirical research substantially influenced international fiscal governance and austerity policy debates. Beyond immediate reputational damage, recent evidence suggests that problematic publications may remain active within the scientific ecosystem for extended periods before retraction. Joëts & Mignon (2026) show that the so-called "zombie papers" (i.e., articles that should be scientifically dead but continue to exert epistemic influence[1]) often survive for years before being formally withdrawn, allowing questionable findings to continue diffusing through citations, policy reports, and subsequent research long after publication and, in some cases, even after retraction notices are issued (Davis (2012), Azoulay et al. (2015), Hsiao & Schneider (2022), and Van Noorden (2023)).

This has fueled broader concerns regarding the reproducibility and governance of modern science, with influential contributions arguing that a substantial share of published findings may be false, fragile, or overstated, thereby calling for deep structural reforms in scientific evaluation and publication systems (Fišar et al. (2024), Pérignon et al. (2024), Ioannidis et al. (2025)). Existing editorial and peer-review mechanisms remain largely reactive rather than preventive. They often detect problematic research only after publication, rely heavily on human reviewers operating under severe informational and time constraints, lack scalable *ex ante* screening infrastructures, and remain vulnerable to strategic adaptation by authors seeking to circumvent detection mechanisms (Wager & Williams (2011), Horbach & Halffman (2019), Heesen & Bright (2021)). More broadly, existing scientific governance mechanisms such as post-publication corrections, plagiarism-detection software, replication initiatives, and formal retraction procedures remain fragmented, costly, and largely non-adaptive, often intervening only after questionable findings have already diffused throughout the scientific ecosystem.

These challenges have become even more acute with the rapid diffusion of generative artificial intelligence (AI) tools, which create new opportunities for the large-scale production, ma-

[1] See Teixeira da Silva & Bornemann-Cimenti (2017) and Joëts & Mignon (2026).

nipulation, and stylistic enhancement of scientific content (Korinek (2023), Korinek (2025), Kusumegi et al. (2025), Li et al. (2025)). In such an environment, editorial systems originally designed for traditional forms of misconduct may prove increasingly inadequate against adversarial AI-assisted behaviors capable of generating more sophisticated and difficult-to-detect forms of manipulation (Stokel-Walker (2022), Holly (2023), Van Dis et al. (2023)).

Editors and reviewers cannot perfectly observe the credibility, robustness, or integrity of scientific outputs prior to publication. Scientific publishing therefore operates under substantial information asymmetries, where authors possess superior information regarding the underlying quality and reliability of their research compared to editors, reviewers, and readers (Le Maux et al. (2019)). Similar informational frictions have long been documented in economics and finance, particularly in credit and insurance markets, where borrowers or insured agents typically hold more information about their underlying risk profiles than lenders or insurers do (Akerlof (1970), Spence (1973), Stiglitz & Weiss (1981)). In such environments, decision-makers must rely on imperfect but observable signals to infer latent risk. In the context of scientific publishing, these signals may include peer-review evaluations, institutional reputation, citation patterns, publication history, and linguistic or stylistic characteristics embedded in scientific texts. Yet, as recent scientific scandals and AI-assisted manipulation practices increasingly demonstrate, many of these traditional signals may be noisy, strategically manipulated, or insufficient to reliably identify problematic research before diffusion occurs (Li et al. (2025)). Confronted with governance challenges structurally similar to those observed in financial screening environments, scientific publishers and journals may therefore require more adaptive and probabilistic credibility assessment systems capable of identifying latent integrity risks from observable textual and publication-related signals.

In response to these challenges, an emerging literature has developed computational approaches aimed at identifying problematic, manipulated, or potentially fraudulent scientific publications. Broadly, three main streams of approaches can be distinguished (see Table A.1 in Appendix A.1 for a detailed overview). First, statistical anomaly detection methods seek to uncover irregularities in numerical outputs and reported data. These approaches include the use of Benford's Law to identify fabricated or manipulated datasets in academic articles (Horton et al. (2020)), the detection of statistical distribution irregularities in clinical trial randomization (Carlisle (2012)), as well as consistency checks based on Likert-scale structures and impossible statistical values in psychology and behavioral sciences (Brown & Heathers (2017), Heathers et al. (2018)). Second, a textual and linguistic stream relies on natural language processing (NLP) and stylometric techniques to detect suspicious linguistic patterns embedded in scientific manuscripts. This literature includes phrase-dictionary approaches designed to identify "tortured phrases", namely artificial paraphrasing patterns often associated with manipulated or automatically generated manuscripts (Cabanac et al. (2021), Cabanac & Labbé (2021), Martel et al. (2023)), together with NLP and stylometric fraud-detection methods focusing on deception signals, semantic inconsistencies, writing style, and authorship patterns (Afroz et al. (2012), Feng et al. (2012), Markowitz & Hancock

(2016)). Third, network and structural approaches examine anomalies in the organization of scientific ecosystems themselves, including irregular co-authorship structures associated with paper mills and authorship-for-sale enterprises (Porter & McIntosh (2024)), citation cartels, and peer-review manipulation networks (Fister & Perc (2016)). While these approaches have substantially advanced the computational detection of scientific misconduct, several important limitations remain. Most existing methods are static and insufficiently adaptive to evolving scientific ecosystems, disciplinary norms, and field-specific publication practices. Moreover, they largely operate *ex post*, thereby limiting their usefulness as early-warning governance mechanisms capable of identifying problematic research before large-scale diffusion occurs. Existing approaches also tend to lack an integrated theoretical framework linking credibility assessment, information asymmetry, and dynamic scientific governance. In addition, the literature remains highly fragmented, with most methods focusing on a single dimension of misconduct such as numerical anomalies, linguistic irregularities, or network structures in isolation. Yet, in rapidly evolving and increasingly AI-assisted scientific environments, effective integrity monitoring likely requires adaptive, theoretically grounded, and multidimensional systems capable of continuously updating credibility assessments from heterogeneous signals over time.

To overcome these limitations, we propose a new theoretical and empirical framework for the probabilistic assessment of scientific credibility and retraction risk. Drawing inspiration from the credit-scoring literature in finance, we conceptualize scientific publishing as a screening environment characterized by latent integrity risk, strategic behavior, and imperfect information. Similar to lenders evaluating borrower default risk under conditions of information asymmetry, editors and publishers must assess the credibility of scientific manuscripts using imperfect and potentially manipulable observable signals. Building on this analogy, we introduce a scientific credibility scoring framework designed to estimate the latent probability that a publication may eventually become problematic or subject to retraction.

More specifically, we develop the ZICO framework (*Zombie Integrity and Credibility Oversight*), a multidimensional credibility scoring system inspired by FICO-style credit scoring models. Rather than relying on a single indicator of misconduct, ZICO integrates heterogeneous signals extracted from both the semantic structure of scientific texts and the broader organization of scientific production. These signals include linguistic complexity and consistency measures derived from natural language processing techniques, together with publication metadata related to co-authorship networks, collaboration structures, institutional and international diversity, and citation behaviors. Using machine-learning classification methods, ZICO transforms these observable characteristics into a continuous credibility score reflecting the estimated probability of future retraction.

Importantly, our framework explicitly recognizes that scientific misconduct detection operates in an adaptive environment shaped by strategic responses from authors seeking to evade editorial screening mechanisms. As generative AI tools increasingly facilitate semantic ma-

nipulation, fluency enhancement, citation engineering, and automated content production, observable credibility signals themselves may become strategically distorted. We therefore conceptualize scientific governance as a dynamic screening problem characterized by adversarial adaptation, similar to strategic behaviors observed in financial fraud, insurance markets, or cybersecurity environments. To address this issue, we further extend the framework through Bayesian updating and a dynamic version of the score, denoted D-ZICO, capable of continuously recalibrating credibility assessments and editorial thresholds as publication environments evolve over time.

Overall, our framework contributes to the literature in several ways. First, unlike most existing approaches that focus on isolated dimensions of misconduct, ZICO provides an integrated and theoretically grounded multidimensional screening architecture. Second, by explicitly framing scientific publishing as an information-asymmetry and adaptive screening problem analogous to credit risk assessment, the paper introduces a novel conceptual bridge between scientific governance and financial screening theory. Third, the framework is designed not merely as an *ex post* diagnostic detection tool, but as a probabilistic early-warning governance mechanism capable of assisting editorial decision-making before problematic research diffuses throughout the scientific ecosystem. More broadly, the paper contributes to ongoing debates on the future of scientific governance in increasingly AI-assisted research environments.

The remainder of this paper is organized as follows. Section 2 introduces the dataset and the variables used to construct the scientific integrity and credibility score. Section 3 develops the theoretical foundations and governance framework underlying the ZICO methodology. Section 4 presents the empirical strategy and main results. Section 5 extends the analysis by examining adaptive governance mechanisms and introducing the dynamic version of ZICO as an early-warning credibility assessment system. Finally, Section 6 concludes and discusses implications for scientific governance and editorial screening practices.

2 Database and variables: Building the zombie radar

This section presents the database of retracted and non-retracted papers used to develop our zombie detection approach in the economics and finance literature. It examines the characteristics and differences between these papers and introduces the key metrics used in our scoring approach.

2.1 Database

To construct our collection of zombie papers (i.e., retracted papers), we use the Retraction Watch Database (RWD), a comprehensive online repository of retracted scientific publications across various disciplines. Widely utilized in research policy and science dynamics literature (see Xu et al. (2023) and Joëts & Mignon (2026), among others), RWD is accessible via the

Crossref API. A key advantage of RWD is that it records research articles (among other documents) along with various metadata, such as titles, journal names, authors, and institutions, over a long period spanning from 1923 to 2023. This provides a centralized platform that aggregates information from multiple sources (see Xu et al. (2023) and Joëts & Mignon (2026) for further details). From the 25,480 available retracted research articles, we focus on papers published in Economics and Finance, yielding a subset of 2,233 research papers.[2]

Our focus is motivated by several key considerations. First, Economics and Finance operate at the intersection of multiple disciplines, drawing on insights from sociology, psychology, political science, statistics, and data science. This interdisciplinary nature makes them particularly susceptible to diverse methodological approaches, varying data standards, and evolving theoretical paradigms, which can, in turn, influence the likelihood and patterns of retraction. Second, given the significant policy and real-world implications of research in these fields—ranging from financial regulation to economic policymaking—ensuring the integrity and reliability of published findings is critical. Retracted studies in Economics and Finance can have far-reaching consequences, including misinforming policy decisions, affecting financial markets, and shaping public discourse. Finally, while retraction patterns have been extensively studied in biomedical and life sciences (Azoulay et al. (2015)), Economics and Finance remain relatively underexplored in this context. By focusing on these fields, we contribute to a growing but still limited body of literature that examines scientific integrity, replication crises, and the dynamics of knowledge correction within the social sciences.

Since our methodological approach to zombie detection relies on semantic structure, access to the full text is essential for analyzing language complexity and consistency. However, obtaining full texts of retracted papers is particularly challenging, as they are often removed from public access for valid reasons, adding complexity to retraction analysis. To address this limitation, we further narrowed our dataset to include only fully accessible papers, resulting in a final set of 88 retracted papers. To serve as a counterpart for our classification model, we also retrieved non-retracted papers. For each retracted paper published in a given journal, year, volume, and issue, we randomly selected one non-retracted paper, yielding a total of 88 non-retracted papers. This selection process resulted in a final sample of 176 papers.[3] Each paper was manually cleaned by removing tables, figures, equations, author names, institutional affiliations, and footnotes to focus solely on the semantic structure and minimize potential biases. Using the DOI of each paper (both retracted and non-retracted), we extracted various metadata, including title, author names, institutional affiliations, country of origin, and journal name. These metadata were used to construct control variables (see Section 2.2.2).

Table 1 provides a summary of the database. As previously discussed, our dataset consists of 176 papers, evenly split between retracted and non-retracted articles. Among the retracted

[2]This selection also includes papers published in accounting, econometrics, management, and marketing journals. The subset was identified using subject classifications provided by RWD and the Scimago journal classification.

[3]Each record has been cross-verified as retracted or non-retracted using Google Scholar and Web of Science.

papers, 42 unique journals are represented, involving a total of 211 unique authors. The most frequently retracted journal is *Economic Modelling*, with 13 occurrences, while the most frequently retracted author is *Khalid Zaman*, who has had 13 papers retracted. Figures B.1 and B.2 in Appendix B present the top ten retracted journals and authors. Notably, all journals except *TEL* are ranked by the Chartered Association of Business Schools (ABS), ranging from ABS 1 (the lowest category, for *IEJ*) to ABS 4* (the highest category, for *RP*, *JoM*, and *AER*). This ranking distribution highlights that retractions occur across various journal tiers, affecting both lower-ranked and top-tier publications. Among the most frequently retracted authors, two cases, *Khalid Zaman* and *Ulrich Lichtenthaler*, stand out due to the volume of their retracted work and their academic influence. *Khalid Zaman*, whose retractions span multiple journals, has been at the center of high-profile cases that raised concerns about research integrity in economic modeling and sustainability studies. *Ulrich Lichtenthaler*, previously a highly cited researcher in innovation management, experienced significant retractions that had a profound impact on the field, particularly in top-ranked business and management journals. His case triggered broader discussions on the reliability of empirical research in these domains and raised concerns about editorial oversight and peer review practices (see Cox et al. (2018) and Horton et al. (2020)). These cases illustrate how retractions can shape academic discourse, particularly when they involve prolific researchers or highly ranked journals, leading to increased scrutiny of research practices and institutional responses.

Table 1: Database description

Metrics	**Values**
Total Papers	176
Retracted papers	88
Total unique retracted journals	42
Total unique retracted authors	211
Most common retracted journals	*Economic Modelling* (13)
Most common retracted authors	Khalid Zaman (13)

Note: This table provides a summary of the database characteristics.

2.2 Variables description

To develop our ZICO method, we define a binary (0-1) endogenous variable to distinguish between retracted and non-retracted papers and construct several exogenous variables expected to play a significant role in classifying zombie papers. First, since semantic and linguistic structures are key in scholarly writing and can signal problematic research, we include measures of language complexity and consistency (see Markowitz & Hancock (2014)). Retracted papers often exhibit linguistic anomalies, such as irregular readability patterns and syntactic inconsistencies, which have been linked to questionable research practices. Second, to account for well-known factors influencing zombie paper dynamics, we incorporate variables on co-authorship networks, team collaboration, institutional and geographic diversity, and citation behaviors (see Bar-Ilan & Halevi (2017), Fanelli (2018), Joëts & Mignon (2026)). Large and

diverse collaborations can enhance oversight but may also facilitate misconduct, while citation anomalies, such as excessive self-citation, can indicate integrity concerns.

2.2.1 Natural language processing: Extracting the hidden decay

To capture the full spectrum of semantic and linguistic characteristics that distinguish zombie papers from non-zombie articles, we consider both language complexity and consistency. All measures are reported and described in Table B.2 in Appendix B.

Language complexity is assessed using 65 different metrics grouped into three main categories:

- *Lexical diversity* (15 metrics) capturing vocabulary richness and variation, assessing how diverse or repetitive the word usage is in a text. Higher lexical diversity indicates a more varied vocabulary, while lower diversity suggests repetitive or constrained word choices.
- *Syntactic complexity* (2 metrics) measuring sentence structure complexity by capturing the depth and intricacy of syntactic constructions. Longer, more embedded clauses suggest greater complexity, while simpler sentence structures indicate lower syntactic sophistication.
- *Readability scores* (48 metrics) assessing text accessibility and comprehension difficulty by accounting for word length, sentence structure, and word familiarity. It help determine how challenging a text is for readers with different levels of expertise.

Language consistency is assessed using 3 different metrics, focusing on:

- *Lexical consistency* measuring the stability of word usage across the text, identifying whether the vocabulary is varied or remains relatively uniform.
- *Part-of-speech consistency* examining grammatical structure stability, indicating whether sentence composition varies significantly across the text or remains uniform.
- *Readability consistency* capturing fluctuations in readability across different sections, detecting whether a text maintains a consistent level of complexity or varies significantly.

As shown in Table B.3 in Appendix B, our metrics capture four classes of linguistic and semantic features, each reflecting a distinct dimension, ranging from *Surface Clarity & Language Norms* to *Stylistic Discipline & Rhetorical Control*.

Since some measures may be highly correlated, leading to dimensionality issues given our 176 observations, we apply principal component analysis (PCA) to the lexical diversity and readability score metrics. This factor model approach reduces redundancy and retains only the most informative dimensions.[4] Figure 1 shows the number of principal components retained

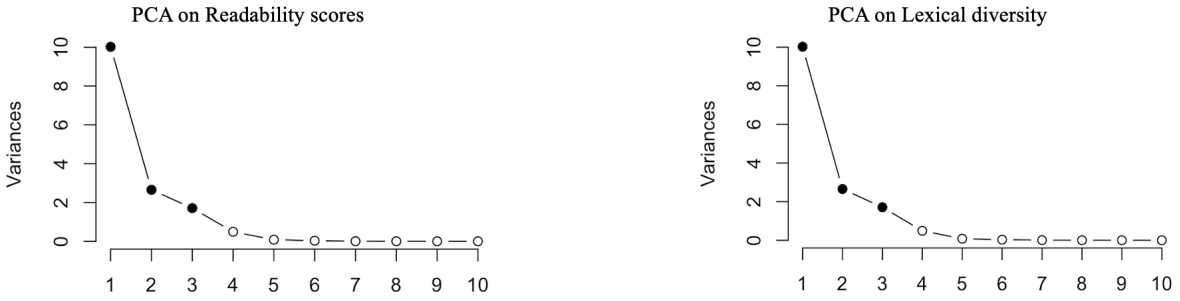
[4]Alternative approaches, such as including all variables, applying PCA to the full dataset, or performing PCA on different subgroups, were tested but resulted in lower predictive performance in the classification experiment. Additional results are available upon request from the authors.

for each score, where black dots indicate the components explaining at least 90% of the total variance. For each metric category, we retain three principal components, further decomposed in Tables 2 and 3.

As shown in Table 2, *Readability scores* capture different dimensions of text complexity. PC1 reflects structural difficulty, primarily driven by *Syllable & word complexity* and *Character & sentence-based readability*, where longer words and sentences increase complexity. PC2 captures syntactic depth, linked to *Sentence & clause complexity* and *Statistical readability & cloze-based readability models*, highlighting embedded clauses and predictability in comprehension. PC3 represents *Word familiarity & list-based metrics*, indicating how well a text aligns with common vocabulary—higher values suggest more specialized or less accessible texts. Table 3 highlights the *Lexical diversity* components. PC1 captures vocabulary growth relative to text length. PC2 measures raw lexical diversity, where higher values indicate a greater proportion of unique words. PC3 introduces a stability factor, capturing fluctuations in lexical variation across text segments.

In summary, readability and lexical diversity capture distinct yet complementary aspects of textual complexity. Readability reflects processing difficulty, influenced by word structure, sentence syntax, and familiarity, while lexical diversity assesses vocabulary variation, considering both uniqueness and distributional stability. These metrics together provide a comprehensive linguistic profile, distinguishing texts by accessibility and lexical sophistication. Combined with the additional metrics detailed in Table B.2 (Appendix B), these variables (11 in total) provide a comprehensive linguistic framework to distinguish zombie papers from non-zombies, capturing critical differences in language complexity, consistency, and semantic structure.

Figure 1: Optimal number of principal components



Note: This figure illustrates the optimal number of principal components for readability scores and lexical diversity. Black dots indicate the components that explain at least 90% of the variance.

Table 2: Factor loadings for readability scores

Principal component	**Classes' names**	**Variables (loading in absolute values)**
PC1	Surface Clarity & Language Norms	Flesch–Kincaid metrics (2.50%); Automated Readability Index / National Reading Index (2.49%); Danielson–Bryan Readability (2.48%).
PC2	Structural Complexity & Cognitive Load	Mean Sentence Length (4.74%); Bormuth Mean Cloze Score (4.59%); Degrees of Reading Power (4.59%).
PC3	Surface Clarity & Language Norms	Dale–Chall Readability (12.23%); Dale–Chall–Powers–Sumner–Kearl Readability (11.70%); Modern Dale–Chall Readability (10.96%).

Note: This table presents the three primary factor loadings for each principal component. Classes' names group variables into families, as reported in Table B.3 in Appendix B. Loadings are expressed in absolute values.

Table 3: Factor loadings for lexical diversity

Principal component	**Classes' names**	**Variables (loading in absolute values)**
PC1	Vocabulary Richness & Density	Maas' a (7.97%); Dugast's Uber Index (7.96%); Maas' logV0 (7.80%).
PC2	Vocabulary Richness & Density	Type–Token Ratio (15.86%); Herdan's C (12.74%); Honoré's statistic (10.54%).
PC3	Stylistic Discipline & Rhetorical Control	Moving-Average TTR (12.39%); Honoré's statistic (11.03%); Yule's K (10.61%).

Note: This table presents the three primary factor loadings for each principal component. Classes' names group variables into families, as reported in Table B.3 in Appendix B. Loadings are expressed in absolute values.

2.2.2 Other variables

In addition to our language complexity and consistency metrics, we construct several variables using the metadata of each paper. These variables capture key elements discussed in the zombie literature (see Fanelli (2009) and Joëts & Mignon (2026)), such as team effects, collaboration structures, and citation patterns, which have been linked to research integrity and retraction likelihood.

To capture the effects of team composition and collaboration dynamics, we construct three variables. First, we measure the number of authors per paper, denoted as NA_i. Large teams can enhance research rigor through increased oversight and expertise, but may also present risks such as diffusion of responsibility or undetected misconduct in multi-authored collaborations (see Foo (2011), Larivière et al. (2015), Aspura et al. (2018), and Zhang et al. (2020)). Second, we incorporate measures of institutional and international diversity (INS_i and INT_i). We define two dummy variables:

$$INS_i = \begin{cases} 1 & \text{if paper } i \text{ has at least one author from another institution} \\ 0 & \text{otherwise.} \end{cases}$$

where INS_i represents institutional diversity for paper i. Similarly,

$$INT_i = \begin{cases} 1 & \text{if paper } i \text{ has at least one author from another country} \\ 0 & \text{otherwise.} \end{cases}$$

where INT_i captures international diversity. Institutional and international diversity are key factors in research evaluation, as multi-institutional and cross-country collaborations tend to receive greater scrutiny and have lower retraction rates, likely due to increased accountability and independent verification of findings (Larivière et al. (2015)). However, certain international collaborations, especially involving predatory publishers, have also been linked to higher retraction risks.

To further refine our classification method, we incorporate two additional metrics related to citation behaviors. First, we measure the total number of cited references for each paper i (CIT_i), serving as an indicator of the paper's engagement with existing literature. Second, we calculate the proportion of self-cited references, which has been flagged as a potential sign of academic misconduct (Sharmaa & Khurana (2025), and Ioannidis et al. (2025)), as:

$$SEP_i = \frac{SE_i}{CIT_i}$$

where SE_i represents self-citations in paper i. These citation metrics, combined with team and collaboration variables, generate five additional exogenous variables, enriching our dataset and strengthening the robustness of our zombie paper classification framework.

While other factors could influence zombie classification, we deliberately exclude certain elements. Specifically, we do not account for prior retraction history (i.e., whether an author had previous retractions before the considered paper), as cases such as *Khalid Zaman* and *Ulrich Lichtenthaler* are relatively isolated, and publication timing complicates systematic evaluation. Additionally, we do not incorporate measures related to data transparency, result consistency, replicability, and reproducibility. While these aspects are crucial in identifying zombification, they suffer from a lack of standardized and recurrent reporting across papers, making objective assessment challenging. Instead, we focus on concrete, measurable elements—metadata and semantic structure—to effectively distinguish zombie papers from non-zombies. Overall, our ZICO detection model incorporates 16 exogenous variables reported in Table B.4 in Appendix B.

3 The ZICO framework: Assessing the integrity and credibility of research

This section introduces the rationale behind the ZICO approach, drawing inspiration from the credit scoring literature in finance, particularly the methodology underlying the FICO scoring system. We first outline the theoretical framework and model before exploring an editorial threshold-based decision rule for distinguishing zombies from non-zombies. Our methodology adopts the perspective of journal editors assessing submissions, aiming to evaluate the risk of a paper being a zombie either before peer review or after the review process.

3.1 From papers to zombies: Theory and model

The core idea of our approach is analogous to credit default risk modeling, where we estimate the probability of default. In our case, however, we model the probability of retraction for a given paper, defining its latent retraction risk Z_i based on a set of characteristics.

We define Z_i as:

$$Z_i = \beta_0 + X_i^{'}\beta + \epsilon_i \tag{1}$$

where Z_i is an unobserved latent retraction risk (higher values indicate a greater likelihood of being a zombie), X_i is the vector of 16 observable paper characteristics discussed in Section 2, β is the vector of estimated coefficients, and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ captures unobservable randomness.

Since retraction is a binary event ($Y_i = 1$ if paper i is retracted, and $Y_i = 0$ otherwise), the choice of a classification model is flexible, provided it is suited for binary outcomes. For illustration, we model it using a logistic function:

$$P(Y_i = 1 \mid X_i) = \frac{1}{1 + e^{-Z_i}} \tag{2}$$

This equation provides the probability that a paper is retracted, given its characteristics. To translate this probability into a credit-like zombie score, we introduce the ZICO measure, defined as:

$$ZICO_i = S_{\min} + (S_{\max} - S_{\min})(1 - P(Y_i = 1 \mid X_i)) \tag{3}$$

where $ZICO_i$ is the zombie credibility score (akin to FICO credit score), S_{min} and S_{max} are the minimum and maximum score values (e.g., 300-850). A higher ZICO score indicates a lower risk of retraction (strong credibility), while a lower ZICO score signals a higher risk of retraction (potential zombie paper).

While we adopt a linear ZICO formulation, Appendix D.1 demonstrates that ZICO remains transformation-invariant across alternative specifications, including log, exponential, square-root, and rank-based transformations.

3.2 Optimal threshold-based editorial decision rule

ZICO assigns each paper a score based on its estimated probability of retraction, with values ranging from 300 to 850, mirroring the FICO credit score scale. Similar to editorial screening, identifying zombie papers requires establishing a classification threshold that separates low-risk (non-zombie) papers from high-risk (zombie) papers. In credit scoring, risk thresholds assess a borrower's creditworthiness; similarly, in the ZICO model, an editorial threshold determines the likelihood of retraction.

The optimal threshold should maximize classification accuracy while minimizing false positives (valid papers wrongly flagged as zombies) and false negatives (actual zombie papers escaping detection). Since retracted papers do not always reveal their flaws immediately, we rely on the ZICO score distribution and use a Finite Mixture Model (FMM) to estimate the latent structure and determine the optimal intersection point, minimizing misclassification. Appendix D.3 provides empirical justification for bimodality.

Inspired by credit scoring, we establish a classification framework:

- Papers below the threshold are classified as "high risk" (potential zombies).
- Papers above the threshold are classified as "safe" (non-zombies).

The true distribution of zombie and non-zombie papers is not directly observable, making threshold determination non-trivial. To address this, FMM identifies latent distributions within ZICO and provides a data-driven threshold estimation.

FMM assumes that the observed ZICO distribution, denoted as $f(ZICO)$, consists of two latent subpopulations, each following a distinct probability distribution:

$$f(ZICO) = \pi_1 f_1(ZICO \mid \theta_1) + \pi_2 f_2(ZICO \mid \theta_2) \tag{4}$$

where $f_1(ZICO \mid \theta_1)$ and $f_2(ZICO \mid \theta_2)$ are the probability density functions corresponding to low-risk (non-zombie) and high-risk (zombie) papers, respectively. π_1 and π_2 are the mixing proportions capturing the relative prevalence of each group ($\pi_1 + \pi_2 = 1$). θ_1 and θ_2 represent the parameters governing each distribution.

Once parameters are estimated via the Expectation-Maximization (EM) algorithm, the optimal threshold $ZICO^*$ is defined at the intersection of the two distributions:

$$f_1(ZICO^* \mid \theta_1) = f_2(ZICO^* \mid \theta_2) \tag{5}$$

This threshold, derived in Appendix D.4, represents the point where a paper is equally likely to belong to either category, ensuring a data-driven classification rule instead of an arbitrary editorial cutoff.[5]

[5] As an alternative, we also consider a three-class classification problem, distinguishing low-, mid-, and high-

4 Empirical results: Tracking the spread of the zombie outbreak

This section presents the empirical results of classifying and scoring zombie and non-zombie papers. First, we estimate the retraction probability for each paper. Second, we compute the ZICO score for both retracted and non-retracted papers. Finally, we determine the optimal threshold point for classification.

4.1 Retraction probability

Following Lessmann et al. (2015) and Hurlin et al. (2024), we evaluated and compared various models to predict the probability of retraction. Our analysis encompasses both interpretable white-box models and more complex black-box models, including logistic regression (LR), LASSO, elastic net (ENet), classification tree (Tree), random forest (RF), XGBoost (XGB), support vector machine (SVM), polynomial SVM, and artificial neural network (ANN). Each model was tuned where necessary and rigorously assessed using a 5-fold cross-validation scheme. Performance was evaluated based on: (i) accuracy (percentage of correctly classified papers), (ii) the area under the curve (AUC), (iii) the false discovery rate (FDR), and (iv) the cost function (CF), following Lessmann et al. (2015) and Hurlin et al. (2024).

The results, presented in Table C.5 and Appendix C, provide a comparative assessment of model performance. While results vary across metrics, XGB emerges as the best-performing model, achieving the highest classification power ($AUC = 0.768$). To further validate this choice, Figure C.3 in Appendix C visualizes ZICO score distributions across models, along with the proportion of retracted papers (in red) and non-retracted papers (in green). The figure reveals that linear parametric models (LR, Lasso, and ENet), standard decision trees, and SVM struggle to effectively separate retracted from non-retracted papers, displaying significant score overlap. This limits their reliability for threshold-based classification. Conversely, ANN and RF exhibit sharp segmentation, but their extreme clustering suggests possible overfitting, potentially reducing generalizability. XGB, however, strikes the best balance, producing a well-defined bimodal structure that minimizes class overlap and enhances classification robustness.

Consequently, the probability function used throughout this paper for classifying zombie and non-zombie papers is:

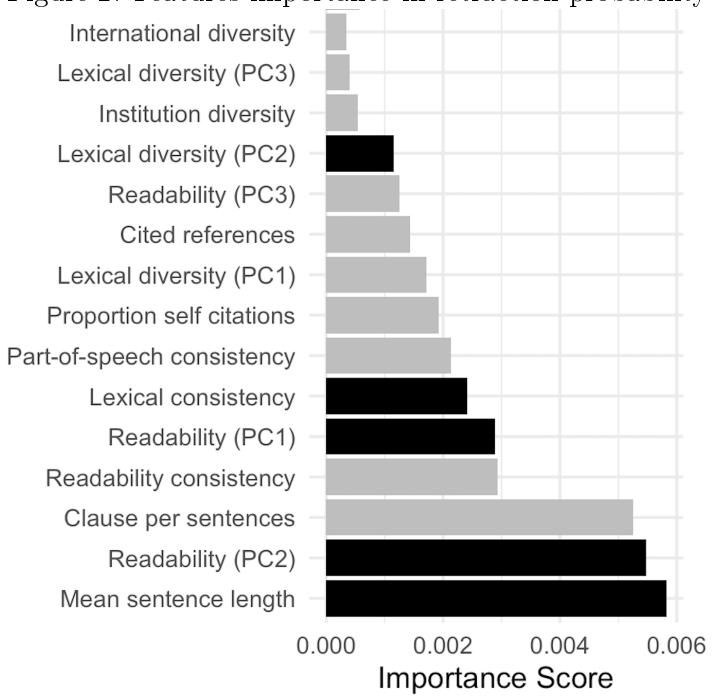
$$P(Y_i = 1 \mid X_i) = f_{XGBoost}(Z_i) \tag{6}$$

where $f_{XGBoost}(Z_i)$ is the estimated probability of retraction obtained from the gradient-boosted model.

risk papers, which results in two intersection points. While this approach introduces additional complexity, results not reported here are available upon request.

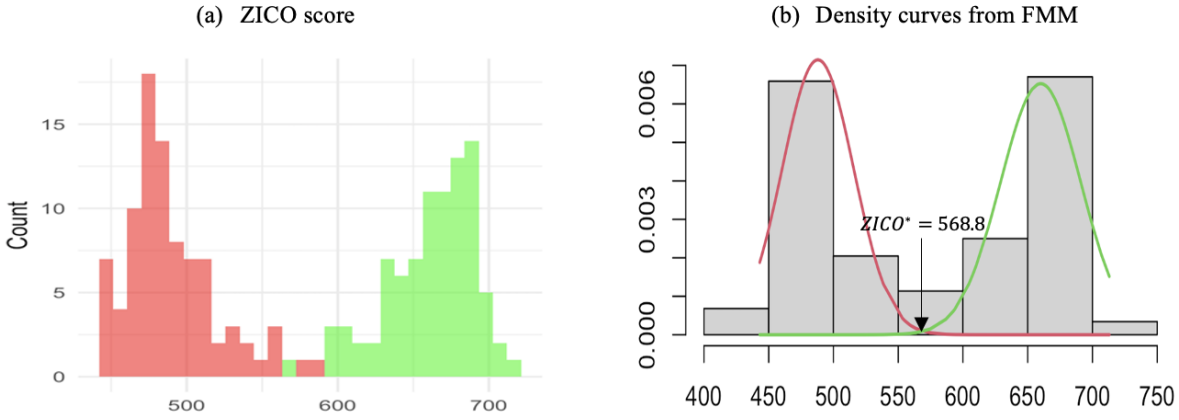
Figure 2 illustrates the absolute importance of each variable in predicting retraction probability using XGB. The directionality of their impact on retraction risk is indicated by color. As can be seen, mean sentence length and Readability PC2 emerge as the most influential variables, suggesting that text complexity, as measured by sentence structure, plays a key role in increasing retraction risk. Similarly, Readability PC1 and Lexical consistency also show a positive association with retraction risk. Conversely, clause per sentence, readability consistency, and part-of-speech consistency are linked to lower retraction risk, indicating that more structured sentence composition and consistent readability contribute to stability in academic papers. Lexical diversity (PC1, PC2, and PC3), along with cited references and proportion of self-citations, exhibits moderate importance, suggesting that variation in vocabulary and citation patterns play a role, though their influence is less pronounced. Interestingly, International diversity and Institutional diversity have relatively low importance, implying that while author diversity may impact retraction risk, linguistic and structural characteristics are the dominant predictive factors. This analysis highlights the critical role of text complexity and consistency in distinguishing papers at higher risk of retraction.

Figure 2: Features importance in retraction probability



Note: This figure reports the absolute feature importance estimated by XGBoost. Bars shown in black correspond to features that increase the risk of retraction, whereas bars shown in grey correspond to features that decrease it.

Figure 3: ZICO score distribution and bimodal classification



Note: This figure presents the distribution of ZICO scores estimated from XGB (panel a) and the corresponding bimodal density curves from FMM (panel b). The optimal threshold, $ZICO^*$, computed as described in Appendix D.4, is indicated at the intersection of the density curves. In red: retracted papers; in green: non-retracted papers.

4.2 ZICO scoring and optimal decision rule

Using the retraction probability estimated from Equation 6, we compute the ZICO score for each paper as defined in Equation 3. Figure 3 presents the distribution of ZICO scores estimated from XGB (panel a) and the corresponding bimodal density curves obtained from FMM (panel b), distinguishing zombie and non-zombie papers. The distribution in panel (a) reveals a clear bimodal structure: retracted papers (in red) tend to have lower ZICO scores, while non-retracted papers (in green) cluster around higher values. This strong separation suggests that ZICO effectively differentiates between high-risk and low-risk papers. In panel (b), the FMM-derived density curves further illustrate this distinction, with the red and green Gaussian distributions representing retracted and non-retracted papers, respectively. The optimal threshold, $ZICO^* = 568.8$, marked at the intersection of these density curves, serves as a classification boundary: papers scoring below this threshold are more likely to be retracted, while those above it are less likely to face retraction. The bimodal nature of both the empirical distribution and the FMM model reinforces the reliability of ZICO as a scoring system, providing a data-driven decision rule for identifying zombie papers. Computing the proportion of safe papers (those above the threshold) yields 50.57%, which perfectly aligns with our dataset, where half of the papers are non-retracted.

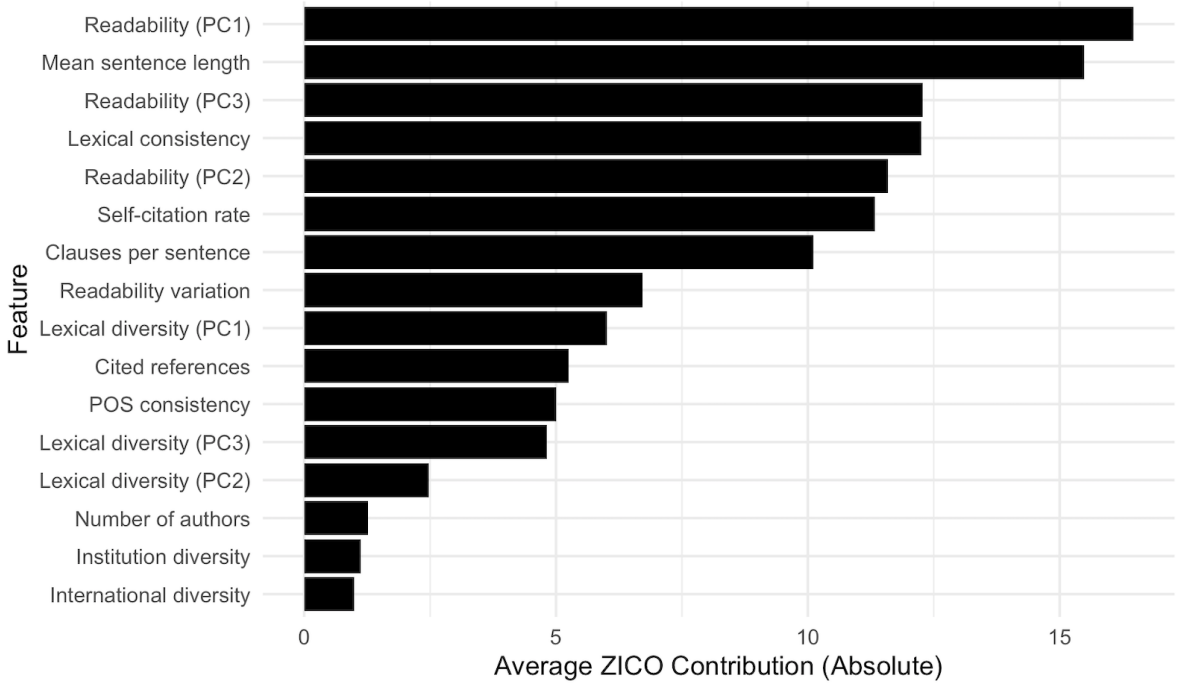
4.3 Explaining ZICO prediction

To better understand how each feature contributes to the ZICO score, we use SHAP (SHapley Additive exPlanations) values derived from the XGBoost model. SHAP offers both global and local explanations of feature contributions by summing the marginal effect of each feature across all possible combinations, providing a consistent and additive decomposition of the

model's prediction.

Figure 4 displays the global SHAP-based contributions to the ZICO score, measured in absolute value. Four distinct blocks of influence emerge from the analysis. First, the most influential predictors—readability PC1 and mean sentence length—each contribute more than 15 points to the ZICO score on average. These variables fall within the readability and syntactic complexity classes, and primarily reflect issues of surface clarity and adherence to language norms, such as the use of complex sentence structures and long words. Second, a cluster of features contributes between 10 and 12 points. This group includes Readability PC2 and PC3, lexical consistency, clauses per sentence, and the proportion of self-citations. These variables span the readability, lexical consistency, and citation behavior classes, and capture aspects of structural complexity, cognitive load, and potential citation manipulation. Third, a moderately influential block (contributing between 5 and 8 points) includes the three components of lexical diversity (PC1, PC2, PC3), cited references, and POS consistency. These indicators reflect vocabulary variation, intertextual integration, and grammatical coherence. Finally, a set of collaboration-related metrics (number of authors, institutional diversity, and international diversity) appear at the lower end of the contribution scale, averaging around 3 points each. While relevant, these team Composition variables exert comparatively less influence on the ZICO score. Overall, this decomposition confirms that ZICO is primarily driven by linguistic and semantic anomalies—especially those related to textual complexity, clarity, and citation practices—rather than by surface-level metadata such as authorship or institutional affiliations. Importantly, the most influential features go beyond simple fluency or proficiency in English. They capture deeper structural and stylistic irregularities in writing that are not merely tied to being a native English speaker, but instead reflect underlying rhetorical coherence, cognitive load, and textual discipline.

Figure 4: Global ZICO contributions

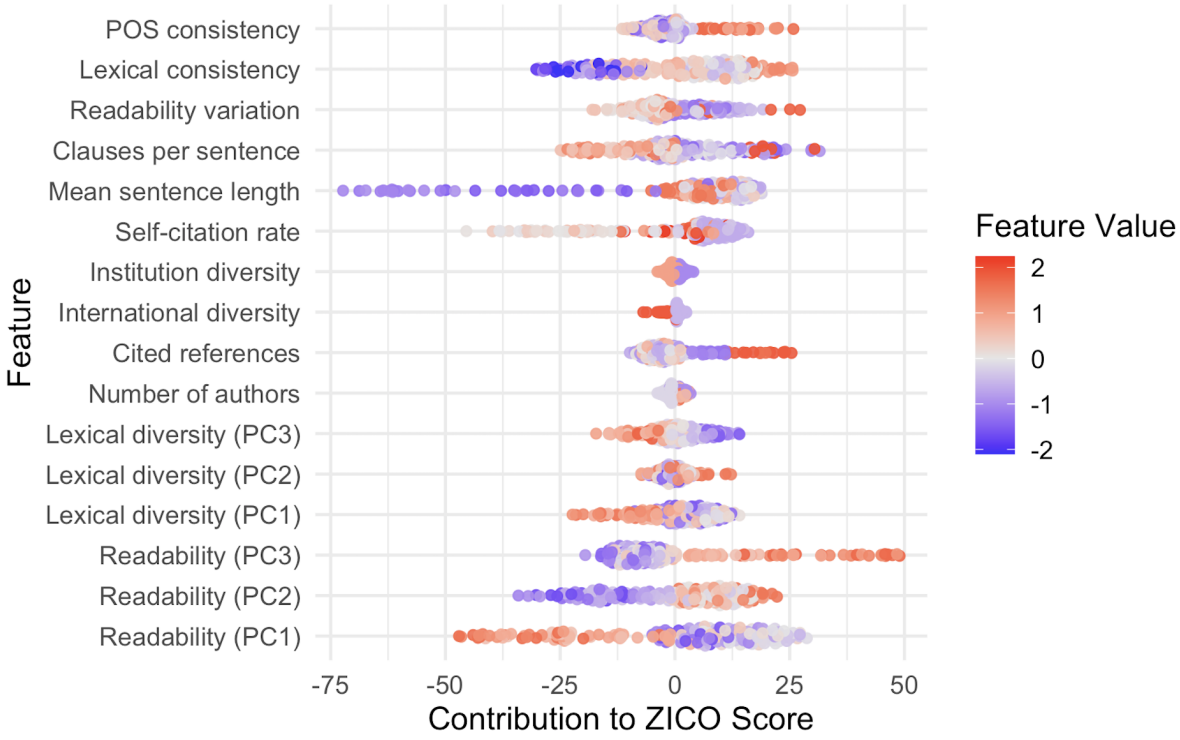


Note: This figure presents the global SHAP-based average absolute contributions of each feature on the ZICO score.

Figure 5 presents the SHAP summary plot for the ZICO score, illustrating the marginal contribution of each feature to the score for all papers in our dataset. Each dot represents a paper, and the color reflects the standardized feature value (Z-score), clipped at the 5th and 95th percentiles to reduce the influence of outliers. Red indicates high feature values, blue low ones, and white or light tones represent values near the mean. Several patterns emerge from this analysis. First, high values of Readability PC1—reflecting structural difficulty linked to longer words and denser sentence constructions—increase the likelihood of retraction (i.e., decrease the ZICO score), suggesting that highly complex writing may be a red flag. Conversely, lower values of Mean Sentence Length (i.e., short, overly simple sentences) are also associated with a decrease in ZICO, reinforcing that both extremes of syntactic complexity may signal problematic writing. Self-citation rate shows a different dynamic: values near the mean are associated with lower ZICO scores (higher retraction risk), while both low and high self-citation levels (blue and red) tend to increase ZICO scores, suggesting a non-linear effect-moderate self-citation may be more suspicious than extremes. For other linguistic dimensions, we observe that low values of Readability PC2 and PC3—capturing syntactic depth and word familiarity—tend to decrease ZICO, while higher values tend to enhance the score. Similarly, low Lexical Consistency, indicating erratic vocabulary usage, also decreases ZICO. Conversely, features such as POS consistency and Cited References display patterns where high values are clearly associated with higher ZICO scores, suggesting that syntactic regularity and strong referencing practices are hallmarks of safer research. In contrast, high values of Lexical Diver-

sity PC1 and PC3, and Clauses per Sentence reduce the ZICO score, suggesting that excessive vocabulary variation or syntactic complexity may also act as warning signals. Overall, the most influential contributions to ZICO scores stem from linguistic complexity and consistency metrics—particularly the components of readability and lexical diversity—underscoring the centrality of semantic structure in differentiating zombie papers from credible ones. These effects are more pronounced than those from metadata features like team size or international collaboration, pointing to language-driven anomalies as a core driver of retraction risk.

Figure 5: SHAP summary plot for ZICO



Note: This figure displays the contribution of each feature to the ZICO score based on its value. For visualization purposes, feature values have been standardized using Z-scores and clipped at the 5th and 95th percentiles to reduce the influence of outliers.

Finally, Figures 6 and 7 present SHAP value breakdowns for a selection of retracted papers authored by K. Zaman and U. Lichtenthaler respectively, alongside non-retracted counterpart articles published in the same journal volumes, issues, and years.

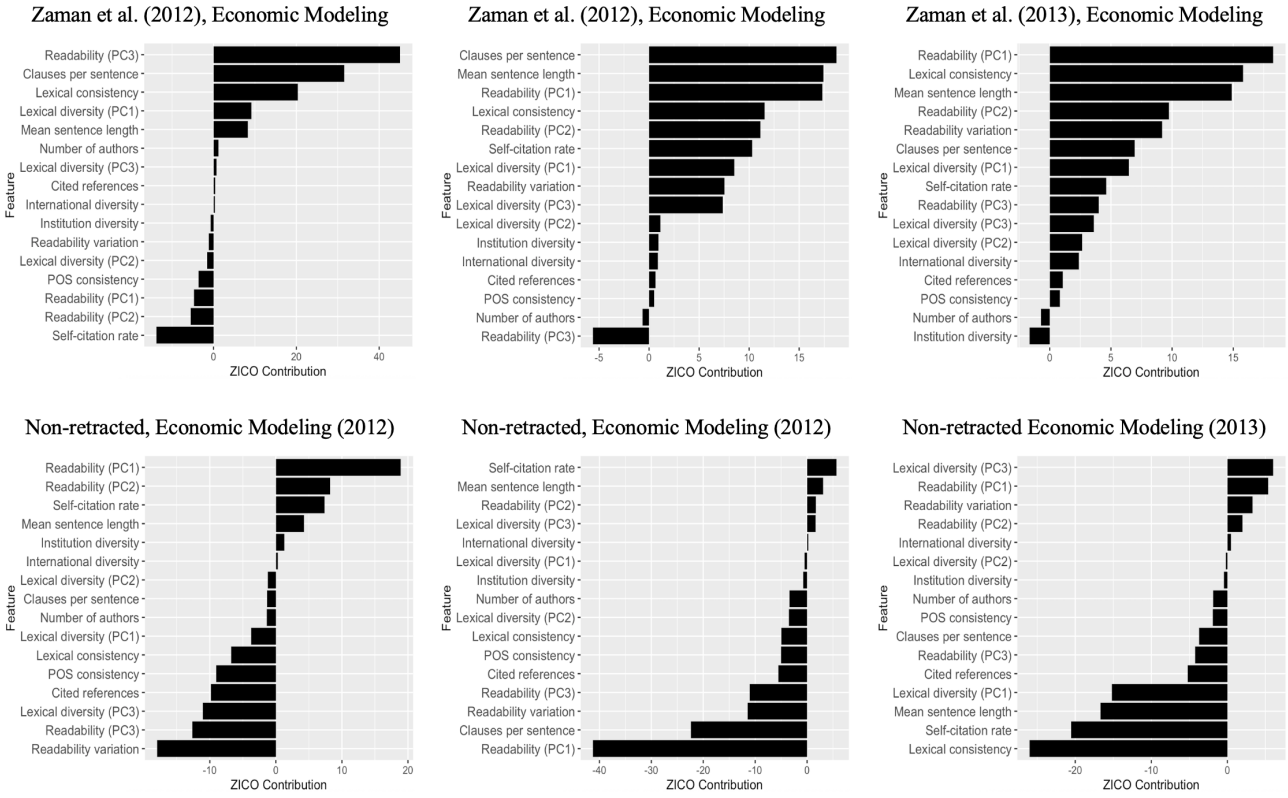
A consistent pattern emerges across Zaman's retracted papers (all published in *Economic Modeling*): linguistic complexity and irregularity are the dominant sources of negative contributions to the ZICO score. In nearly every case, features such as Readability PC1, Lexical Diversity PC1 and PC3, and Clauses per Sentence exert strong negative impacts—often exceeding -10 or even -20 points—indicating that these texts suffer from excessive syntactic complexity, structural inconsistency, or rhetorical disorganization. Additional penalties often

arise from Readability PC2 or Readability Variation, reinforcing the presence of fluency and coherence issues. While certain variables, such as Word Entropy or Self-Citation Rate, may offer modest positive contributions, their influence remains marginal compared to the linguistic disruptions. In contrast, the matched non-retracted papers exhibit a much more balanced SHAP profile. Complexity-related features either contribute positively or remain near neutral, suggesting these papers follow conventional rhetorical and structural norms. Notably, features like Readability PC1 and Lexical Diversity PC3 no longer act as penalties and may instead enhance the ZICO score. Stabilizing indicators—such as POS Consistency, Lexical Consistency, and Cited References—tend to make more substantial positive contributions, underscoring the presence of disciplined and coherent scholarly writing. In short, Zaman's retracted papers are characterized by strong semantic signals of structural disorder and inflated complexity, while the non-retracted control group reflects more typical academic prose, likely reducing their exposure to scrutiny or suspicion.

A similar linguistic signature is observed in Lichtenthaler's retracted papers (from *Technological Forecasting and Social Change*, *Journal of Management Studies*, and *Journal of World Business*), though with subtle distinctions. As in Zaman's case, structural complexity (Readability PC1), vocabulary dispersion (Lexical Diversity PC1/PC3), and citation behavior (notably mid-range Self-Citation Rate) are key drivers of low ZICO scores. However, Lichtenthaler's papers appear to place more weight on Readability PC2 and Readability Variation, indicating that fluctuations in syntactic clarity and flow may be particularly detrimental. By contrast, Lexical Consistency tends to be more penalizing in Zaman's papers than in Lichtenthaler's. The matched non-retracted papers, in turn, show more restrained and coherent SHAP profiles: Readability PC1 and PC2 often have positive or neutral contributions, while Readability PC3 and Clauses per Sentence consistently emerge as top positive predictors, reflecting fluent and structurally sound writing. Self-Citation Rate is generally neutral or only weakly negative in these cases.

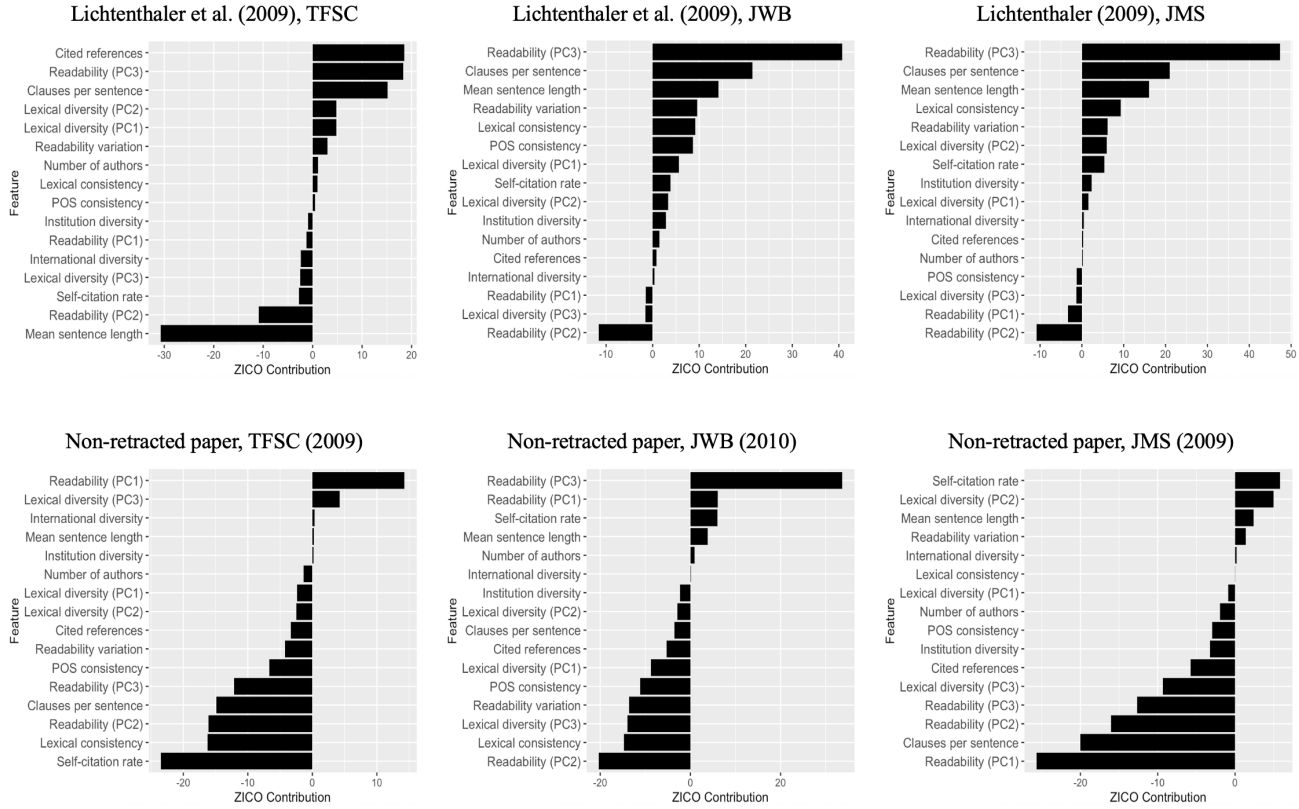
In summary, both Zaman's and Lichtenthaler's retracted papers display pronounced linguistic anomalies, particularly in features related to semantic structure and syntactic complexity, which mark them as zombie candidates. The ZICO framework not only detects these patterns reliably but also pinpoints the specific dimensions—ranging from lexical coherence to citation behavior—that contribute most critically to retraction risk.

Figure 6: SHAP summary for Zaman's papers



Note: This figure presents SHAP-based feature contributions for a selection of retracted papers authored by K. Zaman and coauthors, along with their non-retracted control group counterparts.

Figure 7: SHAP summary for Lichtenthaler's papers



Note: This figure presents SHAP-based feature contributions for a selection of retracted papers authored by U. Lichtenthaler and coauthors, along with their non-retracted control group counterparts.

5 Additional results and extensions

This section explores several extensions. First, we examine the issue of adverse selection and how fraudulent researchers may manipulate their work to conceal misconduct, thereby affecting ZICO. Second, we propose a Bayesian correction method to account for such manipulations. Finally, we extend the static ZICO model to a dynamic version, D-ZICO.

5.1 Dodging Detection: Strategic fraud, adverse selection and bayesian updating

5.1.1 Adverse selection

Detecting scientific fraud or misconduct is crucial to preventing the proliferation of flawed knowledge. However, it remains an inherently challenging task due to the unlimited ways to commit fraud and the limited, non-standardized methods available for detection. The core intuition of our approach is that if fraud is present in a paper, it should leave detectable traces

in the semantic structure of the text.

Unfortunately, much like credit scoring, fraudulent researchers often have the strongest incentives to conceal their misconduct (Necker (2014)), creating an adverse selection dilemma.[6] In academic publishing, fraudulent authors employ various strategies to disguise zombie papers as legitimate research, thereby evading detection by standard editorial and peer-review processes.

Editors face an information asymmetry problem: they cannot observe the true quality of a paper, denoted by ϕ_i, but instead receive an imperfect signal S_i, which depends on observable characteristics and potential manipulations:

$$S_i = \phi_i + \vartheta_i \tag{7}$$

where S_i is the observable credibility signal for paper i, and $\vartheta_i \sim \mathcal{U}(a, b)$ represents strategic noise introduced by authors through manipulations such as semantic alterations, citation distortions, and AI-based text modifications. The uniform distribution assumption reflets that these manipulations have a controlled range rather than an arbitrary spread.

As motivated by the literature, we focus on four key manipulation strategies that fraudulent authors may use to obscure ϑ_i and evade detection (see Table E.9 in Appendix E.1 for a detailed description). Given the growing impact of AI-generated content in academic research (Liyanage et al. (2022), Ma et al. (2023)), we examine three AI-based manipulations and one non-AI manipulation:[7]

1. AI-Generated Text (Ma et al. (2023)): Produces simplified text but increases word repetition, reducing lexical diversity.

2. GPT Fluency Boost (Floridi & Chiriatti (2020), Ma et al. (2023), and Alkaissi & McFarlane (2023)): Enhances text fluency but decreases lexical variety, making detection more difficult.

3. AI Hallucination (Jargon Overuse, Alkaissi & McFarlane (2023)): Introduces excessive jargon and artificial technical sophistication, often making the text seem more complex but potentially unnatural.

4. Citation Stuffing (Fong & Wilhite (2017), Ioannidis (2005)): Artificially inflates credibility by increasing reference counts and self-citations, potentially misleading readers and reviewers regarding the paper's legitimacy.

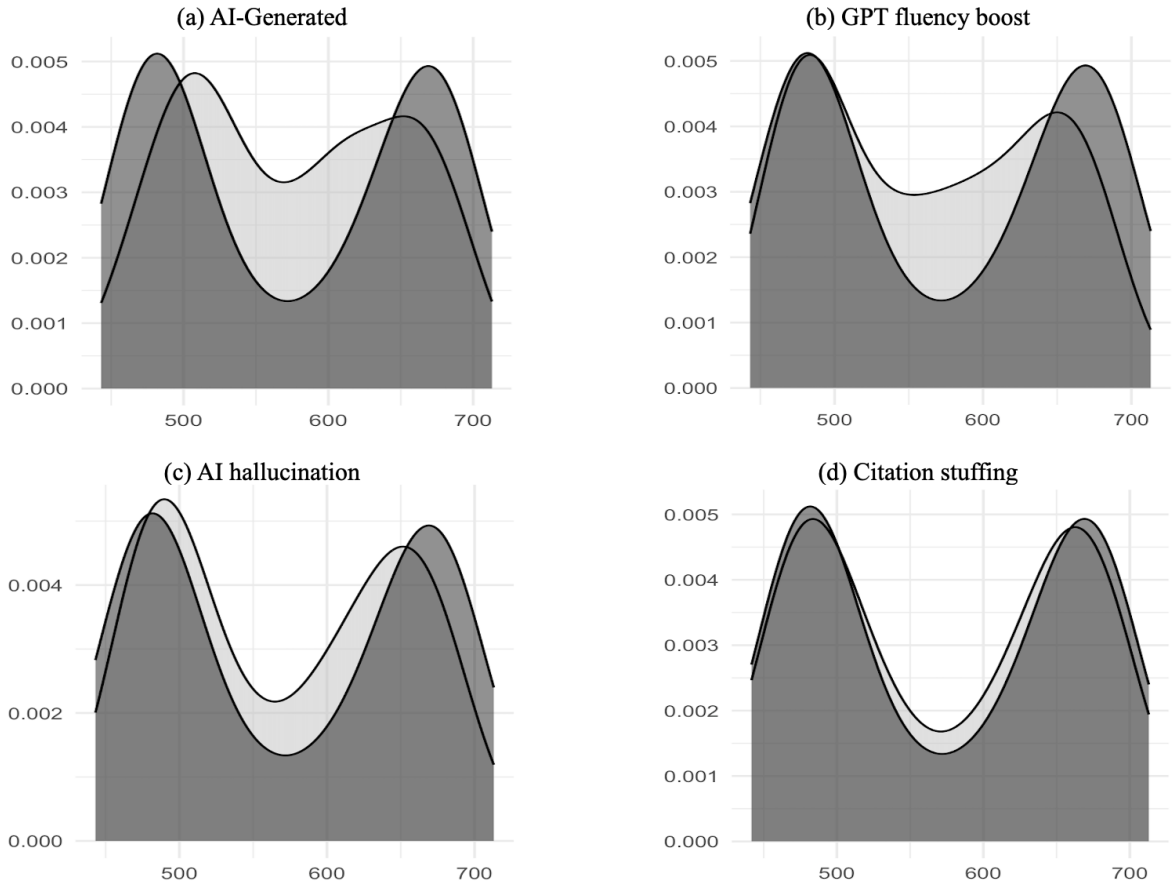
[6] In finance, adverse selection refers to a situation of asymmetric information, where one party in a transaction holds more relevant information than the other, leading to a disproportionate presence of high-risk participants. A classic example is unhealthy individuals purchasing health insurance at higher rates, making it difficult for insurers to assess risk accurately.

[7] Prior studies have also investigated fake collaborations (e.g., ghost authorship and honorary authorship) as a manipulation strategy (Porter & McIntosh (2024)). However, results (not reported but available upon request) indicate that this manipulation does not significantly affect ZICO scoring.

Fraudulent papers employing these manipulations may skew the distribution of ZICO scores and, more importantly, lower the optimal editorial threshold for distinguishing zombies from non-zombies. As a result, the proportion of papers classified as "safe" increases, making it easier for fraudulent research to bypass detection mechanisms (see the mathematical proof in Appendix E.1).

Figure 8 displays the distribution of the ZICO scoring for the original version (in black) versus different manipulations discussed above (in light gray). In all cases, the manipulated distributions deviate from the original ZICO distribution, though to varying extents. AI-based manipulations significantly shift the ZICO distribution, particularly standard AI-generated text (panel a) and GPT fluency boost (panel b). In contrast, the citation stuffing manipulation appears to have only a minor effect on the ZICO distribution. Table 4 reports the optimal editorial threshold and the proportion of safe papers for both original and manipulated ZICO scores. As expected, almost all manipulations lead to a decrease in the editorial threshold compared to the original. The most pronounced changes occur for GPT-based manipulation (a decrease of 53.2 points) and AI hallucination (a decrease of 24.8 points). Consequently, the proportion of falsely classified "safe" papers increases by 11.93% and 4.54%, respectively, leading to more zombie papers being incorrectly labeled as non-zombies. To a lesser extent, standard AI-generated text and citation stuffing also lower the optimal threshold slightly, though their impact is less pronounced.

Figure 8: ZICO distribution after manipulations



Note: This figure illustrates the impact of manipulations on the ZICO distribution. The original ZICO distribution is shown in black, while the manipulated ZICO distributions for each considered manipulation are represented in light gray.

Table 4: Impact of manipulations on the optimal editorial threshold

	Optimal editorial threshold	**Proportion of safe papers (> threshold)**
Original	568.8	50.57%
AI-Generated	562.1	54.55%
GPT fluency boost	515.6	62.50%
AI hallucination	544.0	55.11%
Citation stuffing	565.3	51.14%

Note: This table presents the effects of manipulations on the optimal editorial threshold and the proportion of papers classified as safe.

5.1.2 Bayesian updating

Since manipulations distort the probability of retraction, Bayesian updating is applied to correct the retraction probability estimate, which subsequently updates the ZICO score. We consider several Bayesian updating approaches, including standard and hierarchical methods and a time-based probability adjustment (see Table E.10 in Appendix E.2.1), and evaluate their performances in correcting ZICO using (i) threshold-based AUC; (ii) the Wasserstein distance to the original threshold; and (iii) AUC. As can be seen, from Tables E.12 and E.13 in Appendix E.2.1, the time-based correction is the most effective one, as it assumes that retraction probability changes over time, following a prior updating model

$$P^C(Y_i = 1 \mid S_i, t) = P(S_i \mid Y_i = 1, t).\gamma_t \tag{8}$$

where $P^C(Y_i = 1 \mid S_i, t)$ is the time based corrected probability, $P(S_i \mid Y_i = 1, t)$ the manipulated probability, and γ_i is the time-dependent correction factor (see the proof in Appendix E.2.2).

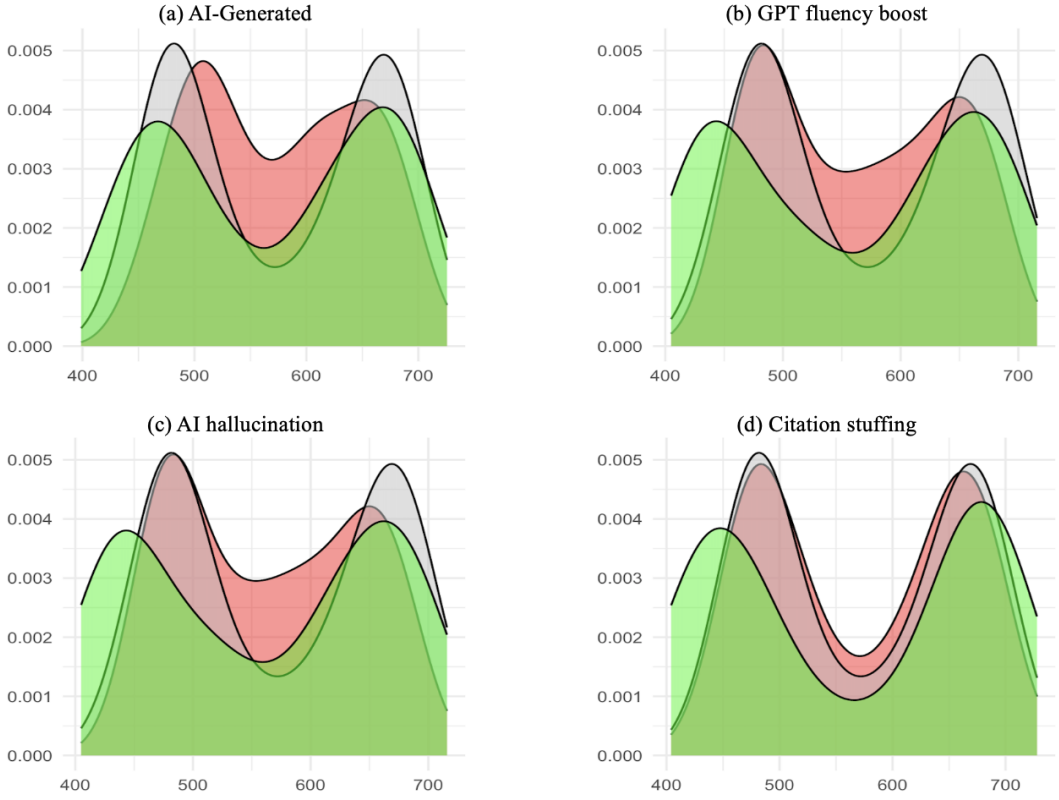
The adjustment factor is assumed to vary within a fixed range of 0.8 to 1.2 across all cases, ensuring that the proportion of zombie papers remains stable over time.[8] This range is chosen to balance stability and adaptability, ensuring that minor fluctuations in retraction probability do not excessively distort ZICO scores while still allowing for dynamic adjustments. Specifically, values below 0.8 tend to overcorrect, increasing the false classification of non-zombies as zombies, whereas values above 1.2 introduce excessive leniency, leading to a higher misclassification of zombies as non-zombies. The proof of the effect of the time-dependent probability adjustment on the optimal threshold and ZICO is available in Appendix E.2.2. Consequently, the time-corrected ZICO score is given by:

$$ZICO_i^{C,t} = S_{\min} + (S_{\max} - S_{\min})(1 - P^C(Y_i = 1 \mid S_i, t)) \tag{9}$$

where $ZICO_i^{C,t}$ is the corrected time-based corrected ZICO.
Figure 9 visualizes the ZICO score distributions under different types of manipulation. The grey density curves represent the original, unmanipulated ZICO distribution, while the red curves depict the manipulated versions, illustrating how various strategies distort the score distributions. The green density curves represent the corrected scores after Bayesian updating, demonstrating how real-time adjustments can effectively mitigate the impact of manipulation. In all cases, Bayesian correction successfully reduces distortions and brings the distributions closer to their original form. While some residual effects remain—particularly in AI hallucination and GPT fluency boost, where the separation between retracted and non-retracted papers is slightly blurred—the correction mechanism significantly improves classification accuracy. These findings underscore adaptive corrections to safeguard the robustness of ZICO scores against strategic behaviors.

[8] We also tested alternative ranges, such as 0.5 to 1.2 and 0.6 to 1.3, which resulted in less effective corrections. Full results are available upon request.

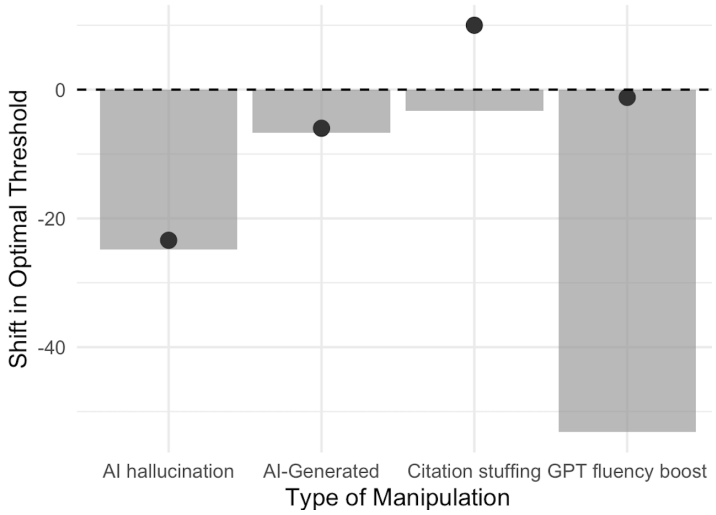
Figure 9: ZICO score manipulations and corrections



Note: This figure presents the original (grey), manipulated (red), and time-based corrected (green) ZICO score distributions.

Figure 10 illustrates the shift in the optimal ZICO threshold resulting from different forms of manipulation. The grey bars indicate the drop in the threshold after manipulation, highlighting how various strategies impact the classification boundary. The black dots represent the corrected thresholds after applying a time-dependent probability adjustment, which helps restore the editorial decision boundary. Notably, AI hallucination and GPT fluency boost cause the most significant downward shifts, increasing the likelihood of manipulated papers evading detection. However, the correction mechanism fully restores the threshold for GPT-based manipulation and partially mitigates the shift for AI hallucination, demonstrating the effectiveness of adaptive Bayesian updating in counteracting certain forms of strategic manipulation.

Figure 10: Threshold shift and correction



Note: This figure shows the shift in the editorial threshold after each manipulation (grey bars). The black dots represent the corrected thresholds obtained after applying a time-dependent probability adjustment.

5.2 D-ZICO: A dynamic editorial early-warning system

A natural extension of the ZICO score is to examine whether such a framework can operate as a real-time editorial decision-support tool capable of identifying potentially problematic manuscripts during the editorial screening process. In practice, editorial environments are inherently dynamic: publication patterns evolve over time, writing conventions change, and the linguistic or structural characteristics associated with future retractions may drift as the scientific ecosystem adapts. Consequently, a purely static classification frontier may become progressively less informative in a changing publication environment.

To address this issue, we introduce a dynamic version of the score, denoted D-ZICO (*Dynamic Zombie Integrity and Credibility Oversight*). Unlike the static ZICO specification, D-ZICO operates within an expanding-window walk-forward framework in which both the predictive model and the resulting classification frontier are recursively updated as new publication data become available. More specifically, at each year t, the predictive model is estimated using only papers published prior to year t, thereby reproducing the informational constraints faced by editors in real time. The estimated model is then used to compute retraction probabilities for papers published in year t, which are subsequently transformed into D-ZICO scores.

Formally, let $\hat{p}_{i,t}$ denote the predicted probability of future retraction for paper i at time t, estimated from a model trained exclusively on observations available before t.[9] The dynamic score is defined as:

[9]For consistency and to ensure comparable model performance, we use the same $f_{XGBoost}$ specification and tuned hyperparameters as in the static ZICO framework.

$$D\text{-}ZICO_{i,t} = S_{\min} + (S_{\max} - S_{\min})\,(1 - \hat{p}_{i,t})\,,$$

where lower values indicate a higher estimated probability of future retraction, while higher values correspond to papers classified as lower-risk. Consistent with the static ZICO framework, the dynamic classification threshold is estimated using a finite mixture model applied to the distribution of D-ZICO scores within the corresponding training sample. However, unlike the static specification, the threshold is itself time-varying:

$$\theta_t = D\text{-}ZICO_t^*$$

where θ_t denotes the optimal dynamic threshold estimated from the distribution of D-ZICO scores computed on papers observed prior to year t. A paper is classified as a potential zombie paper whenever:

$$D\text{-}ZICO_{i,t} < \theta_t.$$

This dynamic specification therefore allows both the predictive structure and the classification frontier to evolve over time, thereby accounting for temporal non-stationarity, concept drift, and changing publication environments. From an editorial perspective, D-ZICO can be interpreted as an adaptive early-warning system that continuously updates its assessment of manuscript credibility as new information accumulates within the scientific literature.

Table 5 reports the global predictive performance and dynamic characteristics of the D-ZICO framework. As can be seen, the dynamic approach achieves a satisfactory global out-of-sample classification performance with an AUC of 0.632, indicating that D-ZICO retains meaningful discriminatory power even in a fully temporal prediction setting. In other words, papers assigned lower D-ZICO scores are substantially more likely to be subsequently retracted than papers assigned higher scores. The model also exhibits a relatively conservative classification profile. Precision reaches 61.1%, implying that a majority of flagged papers are indeed associated with future retractions, while recall remains moderate at 53.2%, indicating that some problematic papers remain undetected. This trade-off is consistent with the intended role of D-ZICO as an early-warning screening mechanism rather than a definitive classification device. Importantly, the model achieves a specificity of 66.1%, suggesting that the procedure limits excessive false-positive detections while still preserving meaningful sensitivity to retraction risk. This point is particularly important in editorial settings, where false accusations may generate substantial investigation costs and reputational concerns for journals and publishers.

The mean D-ZICO score equals 582.3, while the mean dynamic threshold is 569.4, which remains remarkably close to the threshold obtained in the static version of ZICO (568.8). Nevertheless, the threshold evolves over time, ranging from 535.0 to 588.0, reflecting the adaptive nature of the framework and the changing composition of the training population across periods. Such variation is consistent with the objective of D-ZICO, namely continuously recalibrating the classification frontier as new information becomes available.

Overall, the predictive performance of D-ZICO is naturally slightly lower than that of the static ZICO framework. Several factors explain this deterioration. First, D-ZICO relies exclusively on past information when estimating probabilities and thresholds, thereby imposing a stricter and more realistic temporal out-of-sample structure. Second, training sample sizes are substantially smaller during early estimation periods, with the initial estimation window containing only 52 observations. Third, the recursive updating procedure introduces additional estimation noise as both the predictive model and the classification threshold evolve over time. Consequently, D-ZICO should primarily be interpreted as a dynamic early-warning risk indicator designed to support editorial screening and prioritization decisions rather than as a definitive automated classification tool.

Table 5: Global predictive performance and dynamic characteristics of D-ZICO

Metric	**Value**
AUC	0.632
Precision	61.1%
Recall	53.2%
Specificity	66.1%
Mean D-ZICO score	582.3
Mean D-ZICO* (threshold)	569.4
Threshold range	[535.0 ; 588.0]
Initial training sample size	52
Final training sample size	127

Note: D-ZICO is estimated using an expanding-window walk-forward procedure in which both the predictive model and the optimal classification threshold are recursively updated over time. Predictive performance is evaluated in a fully temporal out-of-sample framework.

The usefulness of the dynamic framework is further illustrated in Figures 11, 12, and 13, which respectively display the temporal evolution of the D-ZICO threshold, the global distribution of D-ZICO scores, and the yearly distribution of scores across papers.

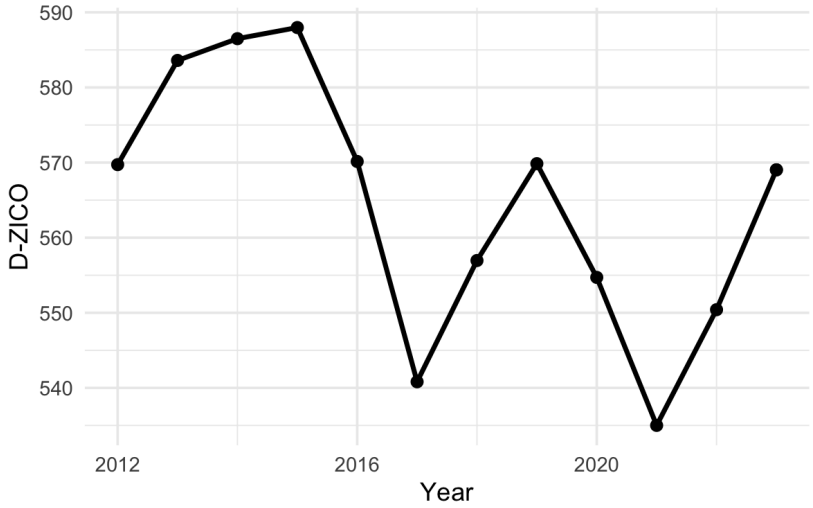
First, Figure 11 highlights the adaptive nature of the D-ZICO framework. The dynamic threshold evolves over time as new papers enter the estimation sample and the underlying distribution of predicted retraction risk changes. Some periods are characterized by a more conservative threshold, particularly around 2021, whereas earlier periods such as 2014-2015 exhibit a relatively less restrictive classification frontier. This temporal variation confirms that D-ZICO continuously recalibrates its screening rule rather than relying on a fixed threshold, thereby adapting to changes in the composition and characteristics of the publication environment.

Second, Figure 12 shows the overall distribution of D-ZICO scores together with the mean

dynamic threshold. Although the distributions of retracted and non-retracted papers partially overlap, retracted papers are substantially more concentrated below the threshold, whereas non-retracted papers tend to exhibit higher D-ZICO scores. This result is consistent with the moderate but meaningful discriminatory power reported previously. At the same time, the figure also illustrates the intentionally conservative nature of the procedure: some papers classified as suspicious are ultimately not retracted, while some retracted papers remain above the threshold. Such a trade-off is expected in an early-warning system designed to prioritize editorial screening while limiting excessive false accusations and unnecessary investigations.

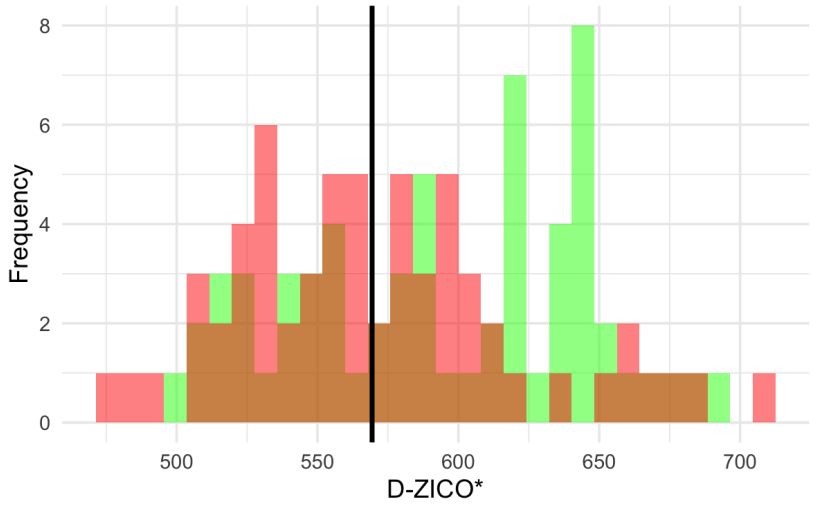
Finally, Figure 13 presents the evolution of individual paper scores over time. Retracted papers generally appear in the lower part of the distribution and cluster more frequently below the dynamic threshold, whereas non-retracted papers are more concentrated at higher score levels. Nevertheless, the figure also confirms that some problematic papers still escape detection, while a limited number of non-retracted papers are flagged as suspicious. This pattern further supports the interpretation of D-ZICO as a probabilistic risk-screening mechanism rather than a deterministic classification device.

Figure 11: D-ZICO time evolution



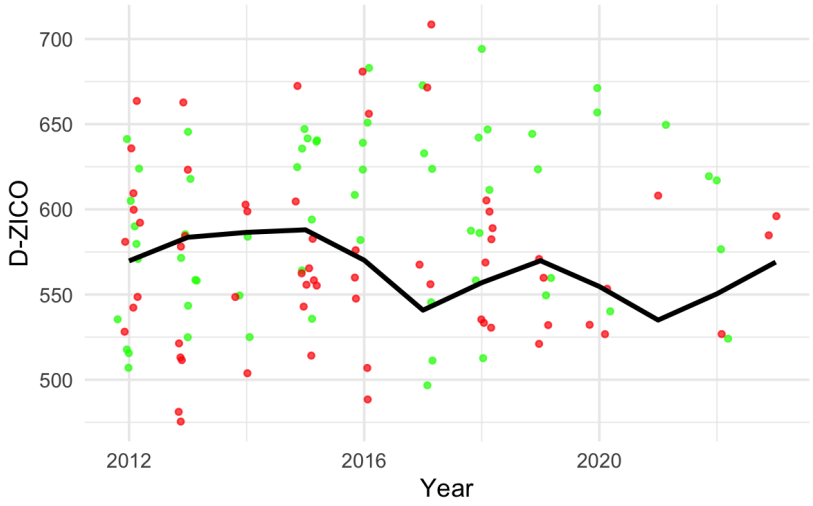
Note: This figure displays the yearly evolution of the dynamic D-ZICO threshold from 2012 to 2023. Variations reflect the recursive updating of the predictive model and the classification frontier as new papers enter the estimation sample.

Figure 12: D-ZICO score distribution



Note: This figure reports the distribution of D-ZICO scores across papers. The black vertical line represents the mean dynamic threshold. Green bars correspond to non-retracted papers, whereas red bars correspond to retracted papers. Lower D-ZICO scores indicate a higher estimated risk of retraction.

Figure 13: D-ZICO score and papers



Note: This figure displays individual D-ZICO scores over time from 2012 to 2023. Green dots correspond to non-retracted papers, whereas red dots correspond to retracted papers. The black line represents the yearly dynamic threshold. Lower D-ZICO scores indicate a higher estimated risk of retraction.

6 Conclusion

This paper develops the ZICO framework, a probabilistic credibility scoring system inspired by financial credit scoring, to estimate the *ex ante* probability that a scientific publication will eventually become problematic or be retracted. By combining linguistic, semantic, and publication-related signals using machine learning, we show that credibility assessment can be framed as an information-asymmetry problem rather than solely as a post-publication problem.

Beyond its methodological contributions, this study raises broader questions about the future organization of scientific governance in an era characterized by rapidly increasing publication volumes, growing methodological complexity, and the widespread adoption of artificial intelligence throughout the research process. Together, these developments are placing increasing pressure on editorial institutions, which must evaluate an expanding body of research with inherently limited human resources. At the same time, AI-assisted writing tools reduce the cost of producing scientifically plausible manuscripts, increasing both the opportunities for scientific discovery and the challenges of quality assurance. In this evolving environment, scientific governance is likely to become more adaptive rather than exclusively reactive. Instead of relying solely on post-publication corrections or binary editorial decisions, credibility assessment may progressively evolve toward continuous, probabilistic evaluation throughout the publication lifecycle. Similar to financial institutions that rely on dynamic credit scores to manage uncertainty rather than predict default with certainty, scientific institutions may increasingly benefit from credibility indicators that quantify risk and are continuously updated as new information becomes available. Such systems should not be viewed as substitutes for scientific judgment, but as complementary mechanisms for improving the allocation of editorial attention and research integrity resources.

These developments also raise questions about the evolution of editorial institutions. Peer review remains the cornerstone of scientific evaluation and provides forms of expert judgment that cannot be fully replicated by algorithmic systems. Nevertheless, the growing complexity and scale of contemporary scientific publishing suggest that peer review alone may no longer be sufficient as a stand-alone governance mechanism. Rather than replacing editors or reviewers, AI-based credibility assessment systems could become one component of a broader, layered governance architecture combining human expertise with complementary computational tools. Within such frameworks, credibility scores could complement existing mechanisms, including plagiarism detection, statistical screening, replication packages, and data availability checks, providing editors with multiple independent sources of information before publication. Importantly, these systems should function as decision-support tools rather than decision-making tools, preserving editorial responsibility while improving efficiency under resource constraints.

At the same time, the deployment of credibility scoring systems raises important ethical and practical considerations. Like any predictive model, ZICO is subject to both false positives

and false negatives. Consequently, credibility scores should never be interpreted as evidence of misconduct, scientific fraud, or future retraction. Instead, they provide probabilistic assessments of latent credibility risk based on observable characteristics that are themselves imperfect proxies. Their responsible use therefore requires transparency, explainability, and continuous validation across disciplines and over time. Credibility assessment should remain assistive rather than deterministic, informing editorial judgment without replacing it. Human oversight remains essential both to interpret algorithmic outputs and to safeguard authors against inappropriate reputational consequences arising from automated screening.

Several avenues for future research emerge from this work. First, future credibility assessment systems could integrate multimodal information beyond textual characteristics, including figures, anomalies in statistical outputs, code repositories, reference trajectories, and peer-review reports. Second, field-specific calibration could improve performance by accounting for disciplinary differences in publication practices, writing conventions, and retraction dynamics. Third, credibility scoring systems could be integrated directly into editorial management platforms, allowing for continuous rather than static assessment throughout the publication process as additional information becomes available. More broadly, this work opens the possibility of developing adaptive scientific governance architectures in which human expertise and artificial intelligence jointly contribute to strengthening research integrity in an increasingly complex scientific ecosystem.

Ultimately, the objective of credibility scoring is not to determine whether a scientific article is true or false. Scientific knowledge advances through cumulative evidence, replication, and critical scrutiny rather than certainty at the moment of publication. Instead, our probabilistic credibility assessment offers a way to quantify uncertainty and prioritize limited editorial resources. In this sense, future scientific governance may increasingly rely on probabilistic credibility indicators to support transparency, accountability, and adaptive oversight while preserving scientific judgment at the core of the evaluation process.

References

Afroz, S., Brennan, M. & Greenstadt, R. (2012), Detecting hoaxes, frauds, and deception in writing style online, *in* '2012 IEEE symposium on security and privacy', IEEE, pp. 461–475.

Akerlof, G. A. (1970), 'The market for "lemons": Quality uncertainty and the market mechanism', *The Quarterly Journal of Economics* **84**(3), 488–500.

Alkaissi, H. & McFarlane, S. I. (2023), 'Artificial hallucinations in ChatGPT: implications in scientific writing', *Cureus* **15**(2).

Aspura, M. Y. I., Noorhidawati, A. & Abrizah, A. (2018), 'An analysis of Malaysian retracted papers: Misconduct or mistakes?', *Scientometrics* **115**, 1315–1328.

Azoulay, P., Furman, J. L., Krieger & Murray, F. (2015), 'Retractions', *Review of Economics and Statistics* **97**(5), 1118–1136.

Bar-Ilan, J. & Halevi, G. (2017), 'Post retraction citations in context: a case study', *Scientometrics* **113**(1), 547–565.

Brown, N. J. & Heathers, J. A. (2017), 'The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology', *Social Psychological and Personality Science* **8**(4), 363–369.

Cabanac, G. & Labbé, C. (2021), 'Prevalence of nonsensical algorithmically generated papers in the scientific literature', *Journal of the Association for Information Science and Technology* **72**(12), 1461–1476.

Cabanac, G., Labbé, C. & Magazinov, A. (2021), 'Tortured phrases: A dubious writing style emerging in science. Evidence of critical issues affecting established journals'. arXiv preprint arXiv:2107.06751.

Carlisle, J. (2012), 'The analysis of 168 randomised controlled trials to test data integrity', *Anaesthesia* **67**(5), 521–537.

Cox, A., Craig, R. & Tourish, D. (2018), 'Retraction statements and research malpractice in economics', *Research Policy* **47**(5), 924–935.

Davis, P. M. (2012), 'The persistence of error: a study of retracted articles on the internet and in personal libraries', *Journal of the Medical Library Association: JMLA* **100**(3), 184.

Fanelli, D. (2009), 'How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data', *PLOS One* **4**(5), e5738.

Fanelli, D. (2018), 'Is science really facing a reproducibility crisis, and do we need it to?', *Proceedings of the National Academy of Sciences* **115**(11), 2628–2631.

Feng, S., Banerjee, R. & Choi, Y. (2012), Syntactic stylometry for deception detection, *in* 'Proceedings of the 50th annual meeting of the association for computational linguistics (volume 2: Short papers)', pp. 171–175.

Fišar, M., Greiner, B., Huber, C., Katok, E., Ozkes, A. I. & Collaboration, M. S. R. (2024), 'Reproducibility in management science', *Management Science* **70**(3), 1343–1356.

Fister, I. & Perc, M. (2016), 'Toward the discovery of citation cartels in citation networks', *Frontiers in Physics* **4**, 240569.

Floridi, L. & Chiriatti, M. (2020), 'Gpt-3: Its nature, scope, limits, and consequences', *Minds and Machines* **30**, 681–694.

Fong, E. A. & Wilhite, A. W. (2017), 'Authorship and citation manipulation in academic research', *PLOS One* **12**(12), e0187394.

Foo, J. Y. A. (2011), 'A retrospective analysis of the trend of retracted publications in the field of biomedical and life sciences', *Science and engineering ethics* **17**, 459–468.

Grieneisen, M. L. & Zhang, M. (2012), 'A comprehensive survey of retracted articles from the scholarly literature', *PLOS One* **7**(10), e44118.

Heathers, J. A., Anaya, J., van der Zee, T. & Brown, N. J. (2018), Recovering data from summary statistics: Sample parameter reconstruction via iterative techniques (sprite), Technical report, PeerJ Preprints.

Heesen, R. & Bright, L. K. (2021), 'Is peer review a good idea?', *The British Journal for the Philosophy of Science* .

Hesselmann, F., Graf, V., Schmidt, M. & Reinhart, M. (2017), 'The visibility of scientific misconduct: A review of the literature on retracted journal articles', *Current sociology* **65**(6), 814–845.

Holly, E. (2023), 'Abstracts Written by ChatGPT Fool Scientists', *Nature* **613**(7944), 423–423.

Horbach, S. P. & Halffman, W. (2019), 'The ability of different peer review procedures to flag problematic publications', *Scientometrics* **118**(1), 339–373.

Horton, J., Kumar, D. K. & Wood, A. (2020), 'Detecting academic fraud using Benford law: The case of Professor James Hunton', *Research Policy* **49**(8), 104084.

Hsiao, T.-K. & Schneider, J. (2022), 'Continued use of retracted papers: Temporal trends in citations and (lack of) awareness of retractions shown in citation contexts in biomedicine', *Quantitative Science Studies* **2**(4), 1144–1169.

Hurlin, C., Pérignon, C. & Saurin, S. (2024), 'The fairness of credit scoring models', *Management Science* .

Ioannidis, J. P. (2005), 'Why most published research findings are false', *PLOS Medicine* **2**(8), e124.

Ioannidis, J. P., Pezzullo, A. M., Cristiano, A., Boccia, S. & Baas, J. (2025), 'Linking citation and retraction data reveals the demographics of scientific retractions among highly cited authors', *PLOS Biology* **23**(1), e3002999.

Joëts, M. & Mignon, V. (2026), 'Slaying the undead: How long does it take to kill zombie papers?', *Research Policy* **55**(2), 105401.

Korinek, A. (2023), 'Generative AI for economic research: Use cases and implications for economists', *Journal of Economic Literature* **61**(4), 1281–1317.

Korinek, A. (2025), 'AI Agents for Economic Research: August 2025 Update to 'Generative AI for Economic Research: Use Cases and Implications for Economists'', *Journal of Economic Literature* **61**(4).

Kusumegi, K., Yang, X., Ginsparg, P., de Vaan, M., Stuart, T. & Yin, Y. (2025), 'Scientific production in the era of large language models', *Science* **390**(6779), 1240–1243.

Larivière, V., Gingras, Y., Sugimoto, C. R. & Tsou, A. (2015), 'Team size matters: Collaboration and scientific impact since 1900', *Journal of the Association for Information Science and Technology* **66**(7), 1323–1332.

Le Maux, B., Necker, S. & Rocaboy, Y. (2019), 'Cheat or perish? A theory of scientific customs', *Research Policy* **48**(9), 103792.

Lessmann, S., Baesens, B., Seow, H.-V. & Thomas, L. C. (2015), 'Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research', *European Journal of Operational Research* **247**(1), 124–136.

Li, C., Hu, X., Xu, M., Li, K., Zhang, Y. & Cheng, X. (2025), Can large language models be trusted paper reviewers? a feasibility study, *in* '2025 IEEE 22nd International Conference on Mobile Ad-Hoc and Smart Systems (MASS)', IEEE, pp. 531–536.

Liyanage, V., Buscaldi, D. & Nazarenko, A. (2022), 'A benchmark corpus for the detection of automatically generated text in academic publications'. arXiv preprint arXiv:2202.02013.

Ma, Y., Liu, J., Yi, F., Cheng, Q., Huang, Y., Lu, W. & Liu, X. (2023), 'AI vs. Human: Differentiation Analysis of Scientific Content Generation'. arXiv preprint arXiv:2301.10416.

Markowitz, D. M. & Hancock, J. T. (2014), 'Linguistic traces of a scientific fraud: The case of Diederik Stapel', *PLOS One* **9**(8), e105937.

Markowitz, D. M. & Hancock, J. T. (2016), 'Linguistic obfuscation in fraudulent science', *Journal of Language and Social Psychology* **35**(4), 435–445.

Martel, E., Lentschat, M. & Labbé, C. (2023), Detection of tortured phrases in scientific literature, *in* 'Proceedings of the Second Workshop on Information Extraction from Scientific Publications', pp. 43–48.

Necker, S. (2014), 'Scientific misbehavior in economics', *Research Policy* **43**(10), 1747–1759.

Oransky, I. (2022), 'Retractions Are Increasing, but Not Enough', *Nature* **608**(7921), 9.

Pérignon, C., Akmansoy, O., Hurlin, C., Dreber, A., Holzmeister, F., Huber, J., Johannesson, M., Kirchler, M., Menkveld, A. J., Razen, M. et al. (2024), 'Computational reproducibility in finance: Evidence from 1,000 tests', *The Review of Financial Studies* **37**(11), 3558–3593.

Porter, S. J. & McIntosh, L. D. (2024), 'Identifying fabricated networks within authorship-for-sale enterprises', *Scientific Reports* **14**(1), 29569.

Sharmaa, K. & Khurana, P. (2025), 'Retracted Citations and Self-citations in Retracted Publications: A Comparative Study of Plagiarism and Fake Peer Review'. arXiv preprint arXiv:2502.00673.

Spence, M. (1973), 'Job market signaling', *The Quarterly Journal of Economics* **87**(3), 355–374.

Steen, R. G., Casadevall, A. & Fang, F. C. (2016), Why has the number of scientific retractions increased?, *in* A. E. Kazdin, ed., 'Methodological Issues and Strategies in Clinical Research', 4 edn, American Psychological Association, pp. 557–568.

Stiglitz, J. E. & Weiss, A. (1981), 'Credit rationing in markets with imperfect information', *The American Economic Review* **71**(3), 393–410.

Stokel-Walker, C. (2022), 'AI Bot ChatGPT Writes Smart Essays—Should Professors Worry?', *Nature* .

Teixeira da Silva, J. A. & Bornemann-Cimenti, H. (2017), 'Why do some retracted papers continue to be cited?', *Scientometrics* **110**(1), 365–370.

Van Dis, E. A., Bollen, J., Zuidema, W., Van Rooij, R. & Bockting, C. L. (2023), 'ChatGPT: five priorities for research', *Nature* **614**(7947), 224–226.

Van Noorden, R. (2023), 'More Than 10,000 Research Papers Were Retracted in 2023-A New Record', *Nature* **624**, 479–481.

Wager, E. & Williams, P. (2011), 'Why and how do journals retract articles? An analysis of Medline retractions 1988–2008', *Journal of Medical Ethics* **37**(9), 567–570.

Xu, H., Ding, Y., Zhang, C. & Tan, B. C. (2023), 'Too official to be effective: An empirical examination of unofficial information channel and continued use of retracted articles', *Research Policy* **52**(7), 104815.

Zhang, Q., Abraham, J. & Fu, H.-Z. (2020), 'Collaboration and its influence on retraction based on retracted publications during 1978–2017', *Scientometrics* **125**, 213–232.

Appendix

A Literature review

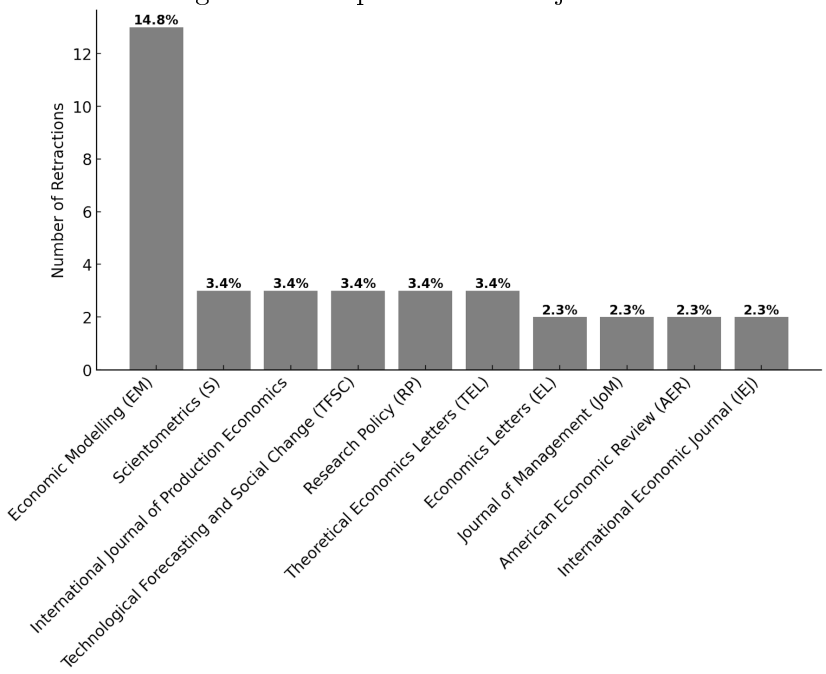
Table A.1: Main computational approaches for detecting problematic scientific publications

Stream	Focus	Typical methods & examples	Main limitations
Statistical anomaly detection	Fabricated or manipulated numerical data	Benford's Law (Horton et al. (2020)); statistical distribution irregularities in clinical trial randomization (Carlisle (2012)); GRIM and SPRITE consistency checks for Likert-scale structures and impossible statistical values (Brown & Heathers (2017), Heathers et al. (2018))	Narrow misconduct coverage; mainly focused on numerical inconsistencies and fabricated data
Textual and linguistic detection	Linguistic irregularities and deceptive writing patterns	Phrase-dictionary approaches for tortured phrases (Cabanac et al. (2021), Cabanac & Labbé (2021), Martel et al. (2023)); NLP and stylometric fraud detection based on deception signals, semantic inconsistencies, writing style, and authorship patterns (Afroz et al. (2012), Feng et al. (2012), Markowitz & Hancock (2016))	Easily adaptable to evolving writing strategies and AI-generated content; often limited to textual dimensions
Network and structural detection	Ecosystem anomalies and collaborative manipulation	Detection of anomalous co-authorship structures associated with paper mills and authorship-for-sale enterprises (Porter & McIntosh (2024)); citation cartels and peer-review manipulation networks (Fister & Perc (2016))	Still emerging; limited standardization and integration across scientific fields

Note: This table provides an overview of the three main computational approaches to detecting problematic, manipulated, or potentially fraudulent scientific publications.

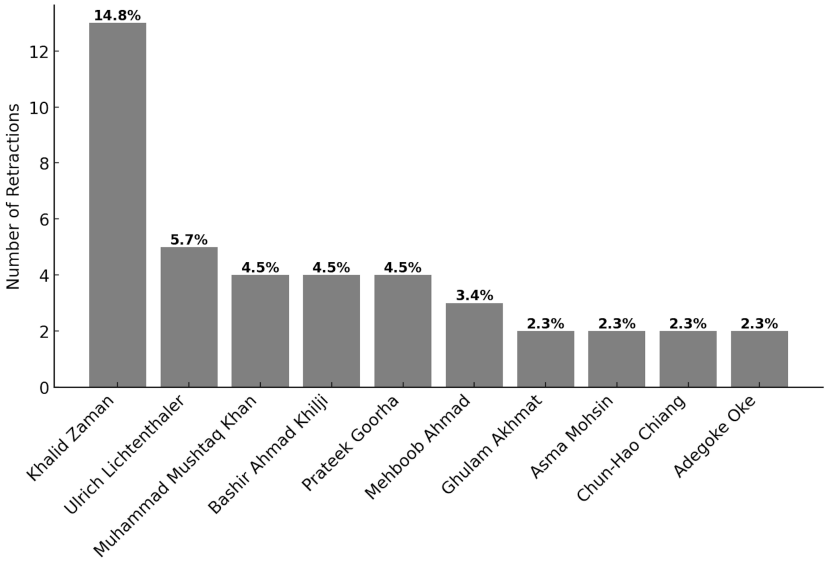
B Database and variables description

Figure B.1: Top ten retracted journals



Note: This figure presents the top ten journals with the highest number of retracted papers, along with their occurrence and proportion of the total retracted papers.

Figure B.2: Top ten retracted authors



Note: This figure presents the top ten authors with the highest number of retracted papers, along with their occurrence and proportion of the total retracted papers.

Table B.2: Language complexity and consistency metrics

	Metrics	Description
Language complexity		
Lexical diversity	Type–Token Ratio (TTR)	Ratio of unique words (types) to total words (tokens).
	Herdan's C	Adjusted TTR logarithmically to account for text length.
	Guiraud's Root TTR	Normalized TTR obtained by dividing the number of types by the square root of the number of tokens.
	Carroll's Corrected TTR	Adjusted TTR obtained by dividing the number of types by $\sqrt{2 \times \text{tokens}}$.
	Dugast's Uber Index	Logarithmic measure adjusting for text-length effects.
	Summer's Index	Uses double logarithms to smooth variability.
	Yule's K	Measures lexical concentration.
	Honoré's statistic	Penalizes frequently used words to estimate vocabulary richness.
	Simpson's D	Measures the probability that two randomly selected words are the same.
	Herdan's Vm	Averages TTR over overlapping text segments.
	Maas' a	Logarithmic measure balancing diversity and text length.
	Maas' logV0	Log-transformed estimate of vocabulary growth.
	Maas' lnV0	Extended version of logV0, adjusting for variations in text length.
	Moving-Average TTR	TTR computed across a moving window to assess text stability.
	Mean Segmental TTR	TTR averaged across equal-length text segments.
Syntactic complexity	**Mean sentence length**	Average number of words per sentence, indicating sentence complexity and verbosity.
	Count clauses per sentence	Average number of clauses per sentence, capturing syntactic depth and subordination.
Readability score	**Character & sentence-based readability scores** (×10)	Estimate readability using characters per word and sentence length.
	Word familiarity & list-based scores (×5)	Compare words against predefined lists of "easy" and "difficult" words.
	Syllable & word complexity-based scores (×5)	Consider syllables per word and sentence length to determine complexity.
	Sentence & clause complexity-based scores (×4)	Focus on sentence structure and clause density.
	Polysyllabic word-based scores (×5)	Count complex words containing three or more syllables to estimate the required education level.
	Statistical & cloze-based readability models (×7)	Predict comprehension using statistical models or cloze-probability tests.
	Alternative complexity measures (×12)	Apply non-traditional formulas, probabilistic models, or scoring systems.
Language consistency		
Lexical consistency	**Word entropy**	Captures lexical diversity by assessing how unpredictably words appear in the text.
Part-of-speech consistency	**POS entropy**	Measures variability in part-of-speech distributions, reflecting syntactic diversity.
Readability consistency	**Readability standard deviation**	Quantifies variability in readability scores across different measures.

Note: This table presents language complexity metrics categorized into lexical diversity, syntactic complexity, and readability scores. Additionally, language diversity is further divided into lexical consistency, part-of-speech consistency, and readability consistency. Color classification is discussed in Table B.3 in Appendix B.

Table B.3: Metrics' groups

Classes' names	**What they capture**	**Metrics' names**
Surface Clarity & Language Norms (in blue)	Sentence fluency, lexical familiarity, and grammatical coherence associated with "good English".	Mean sentence length; count of clauses per sentence; character- and sentence-based readability scores; word-familiarity and list-based scores; syllable- and word-complexity scores; sentence- and clause-complexity scores.
Vocabulary Richness & Density (in green)	Lexical diversity, word rarity, and vocabulary expansion, reflecting density rather than fluency.	Type–Token Ratio (TTR); Herdan's C; Guiraud's Root TTR; Carroll's Corrected TTR; Dugast's Uber Index; Summer's Index; Yule's K; Honoré's statistic; Simpson's D; Herdan's Vm; Maas' a; Maas' logV0; Maas' lnV0; polysyllabic-word-based scores.
Structural Complexity & Cognitive Load (in black)	Syntactic embedding and processing demands associated with comprehension difficulty rather than poor grammar.	Statistical and cloze-based readability models (e.g., DRP, Bormuth MC, and FORCAST); alternative complexity measures (e.g., SMOG, FOG, RIX, and ELF).
Stylistic Discipline & Rhetorical Control (in purple)	Consistency and control across a document, signaling rhetorical planning rather than fluency.	Moving-Average TTR; Mean Segmental TTR; word entropy; readability standard deviation.

Note: This table provides a quantitative classification of the NLP metrics according to the linguistic properties they capture.

Table B.4: Final set of variables

	Variables' names
Language complexity	
Lexical diversity	LPC1 (first principal component) LPC2 (second principal component) LPC3 (third principal component)
Syntactic complexity	MML CPS
Readability score	RPC1 (first principal component) RPC2 (second principal component) RPC3 (third principal component)
Language consistency	
Lexical consistency	WENT
Part-of-speech consistency	POS
Readability consistency	RSD
Other variables	
Institutional diversity	INS
International diversity	INT
Cited references	CIT
Proportion of self-citations	SEP
Number of authors	NA

Note: This table presents the final set of variables considered in the analysis.

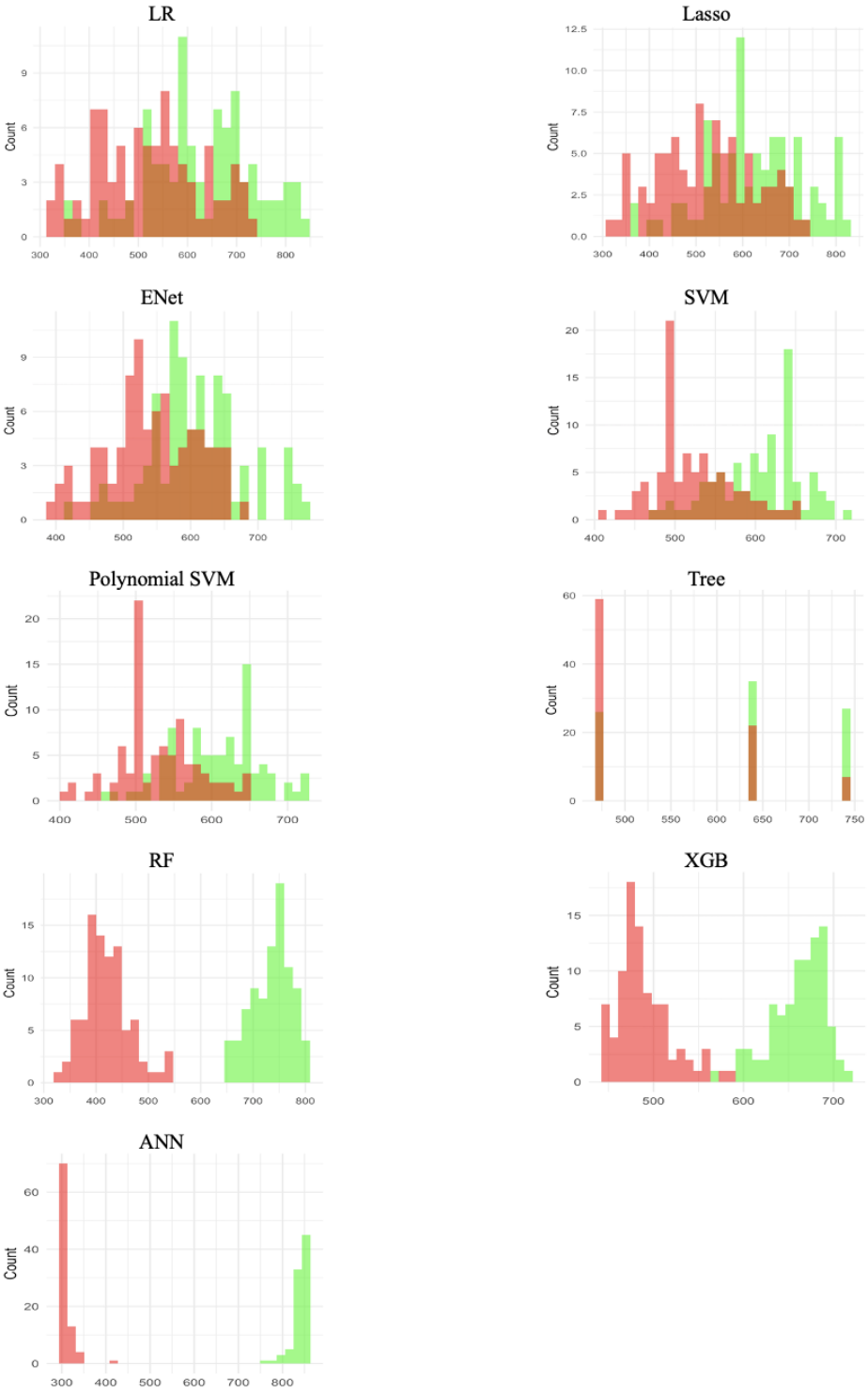
C Models' performance and scoring

Table C.5: Models' performances

Models	Accuracy	AUC	FDR	CF
LR	0.585	0.641	0.431	1080
Lasso	0.471	0.642	0.534	1390
ENet	0.517	0.675	0.545	1220
SVM	0.500	0.624	0.522	1300
Polynomial SVM	0.500	0.678	0.477	1300
Tree	0.505	0.649	0.477	1320
RF	0.431	0.718	0.579	1490
XGB	0.488	0.768	0.500	1360
ANN	0.545	0.624	0.397	1250

Note: This table presents the performance of each considered model using where accuracy (percentage of correctly classified papers), the area under the curve (AUC), the false discovery rate (FDR), and the cost function (CF).

Figure C.3: ZICO distributions across models



Note: This figure presents the ZICO score distribution for each estimated model. Non-retracted papers (non-zombies) are shown in green, while retracted papers (zombies) are shown in red.

D ZICO properties and editorial decision rule

This section provides a detailed evaluation of the ZICO score. First, we assess whether ZICO is transformation-invariant. Second, we examine its robustness to different model specifications (model-invariance). Third, we test for bimodality in the ZICO distribution. Finally, we use FMM to determine the optimal editorial decision threshold that distinguishes zombies from non-zombies.

D.1 Is ZICO transformation invariant?

To ensure that the classification and ranking of papers remain stable across different formulations of the ZICO score, we test both theoretically and empirically whether alternative ZICO transformations preserve the relative ordering of papers.

Alongside the linear ZICO transformation used in the main analysis, we consider four alternative formulations:

- Log-transformed ZICO: $ZICO = 300 + 550 \times \frac{(1-\log(1+P(Y=1|X))}{\log 2}$
- Exponential ZICO: $ZICO = 300 + 550 \times (1 - e^{-P(Y=1|X)})$
- Square-root ZICO: $ZICO = 300 + 550 \times (1 - \sqrt{P(Y=1 \mid X)})$
- Rank-based ZICO: $ZICO = 300 + 550 \times \left(1 - \frac{\text{rank}(P(Y=1|X))}{N}\right)$

We first formally prove that both the standard linear ZICO and all alternative transformations are monotonic functions of the probability of retraction, $P(Y = 1 \mid X)$. The standard ZICO formulation is:

$$ZICO_i = 300 + 550 \times (1 - P(Y = 1 \mid X))$$

Differentiating with respect to $P(Y = 1 \mid X)$ gives:

$$\frac{d}{dP(Y = 1 \mid X)} ZICO = -550$$

Since $-550 < 0$, the function is strictly decreasing, meaning that:

$$P(Y = 1 \mid X_i) > P(Y = 1 \mid X_j) \Rightarrow ZICO_i < ZICO_j$$

This ensures that monotonic order is preserved.

We now demonstrate that the alternative transformations also maintain monotonicity. For the logarithmic transformation, the derivative is:

$$\frac{d}{dP(Y = 1 \mid X)} ZICO = -\frac{550}{(1 + P(Y = 1 \mid X) \log 2}$$

Since this derivative is always negative, the transformation is monotonic. The exponential transformation gives:

$$\frac{d}{dP(Y=1 \mid X)} ZICO = -550e^{-P(Y=1|X)}$$

Since $e^{-P(Y=1|X)} > 0$, the function is strictly decreasing and monotonic. The square-root transformation gives:

$$\frac{d}{dP(Y=1 \mid X)} ZICO = -\frac{550}{(2\sqrt{P(Y=1 \mid X)}}$$

which is also monotonic, as is the rank-based transformation, which preserves relative positions by definition.

To further validate empirical consistency, we compute the rank correlation (Spearman's ρ) between different ZICO formulations. Table D.6 reports these correlations, showing that ZICO is transformation-invariant, as rankings remain nearly identical across all functional forms.

Table D.6: ZICO rank correlation across transformations

	Linear ZICO	**Log-transformed ZICO**	**Exponential ZICO**	**Square-root ZICO**	**Rank-based ZICO**
Linear ZICO	1.000	1.000	-0.999	0.999	0.972
Log-transformed ZICO	X	1.000	-1.000	1.000	0.972
Exponential ZICO	X	X	1.000	-1.000	-0.972
Square-root ZICO	X	X	X	1.000	0.973
Rank-based ZICO	X	X	X	X	1.000

Note: This table reports Spearman's ρ rank correlation for each ZICO transformation. High correlation values confirm that rankings are preserved across different ZICO formulations.

D.2 Is ZICO model-invariant?

While XGB is our best-performing model, we assess whether ZICO remains invariant to model specification. Table D.7 reports Spearman's ρ rank correlation of ZICO scores estimated across different models relative to XGB. The results indicate a high degree of invariance across various probability estimation methods, particularly among tree-based and ensemble models.

XGB (our primary model) and RF yield highly correlated ZICO scores, suggesting that these models capture similar underlying patterns. In contrast, regression-based models such as LR, Lasso, and ENet exhibit moderate correlation, reflecting differences in how they characterize retraction risk. SVM and ANN also produce comparable rankings but introduce minor variations. However, the Tree model demonstrates the weakest correlation, indicating greater instability and poorer generalization.

Despite these variations, ZICO preserves rank-order consistency and classification performance across models, confirming its robustness to different probability estimation approaches.

Table D.7: ZICO rank correlation across model specifications

	ZICO XGB
ZICO Logistic	0.521
ZICO Lasso	0.516
ZICO ENet	0.510
ZICO SVM	0.697
ZICO polynomial SVM	0.666
ZICO Tree	0.536
ZICO RF	0.948
ZICO ANN	0.811

Note: This table reports Spearman's ρ rank correlation between ZICO scores computed from different model specifications relative to XGB. High correlation values indicate that rankings are preserved across models.

D.3 Is ZICO bimodal?

As observed in Figure C.3 in Appendix C, the XGB-based ZICO distribution exhibits strong signs of bimodality. This observation is crucial for identifying zombie papers, as it suggests that retracted and non-retracted papers follow two distinct distributions. To formally test this hypothesis, Table D.8 reports a comparison of the Bayesian Information Criterion (BIC) and Raftery's BIC probability between a unimodal and a bimodal specification.[10]

The unimodal distribution assumes a single normal density:

$$f(ZICO) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(ZICO-\mu)^2}{2\sigma^2}}$$

while the bimodal distribution models a mixture of two normal densities:

$$f(ZICO) = \pi_1 \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(ZICO-\mu_1)^2}{2\sigma_1^2}} + \pi_2 \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{(ZICO-\mu_2)^2}{2\sigma_2^2}}$$

where π_1 and π_2 represent the mixing proportions, and (μ_1, σ_1) and (μ_2, σ_2) denote the means and standard deviations of the two underlying distributions.

[10] Please note that in the context of model selection for FMM (particularly using Mclust in the R package), the BIC is interpreted to favor higher values. This approach contrasts with many other statistical contexts where a lower BIC is preferred. Thus, the interpretation shifts to prioritize models that, despite their complexity, provide a significantly better fit to the data. In this scenario, a higher BIC value indicates that the additional complexity, introduced by more mixture components, is justified by a superior fit to the data.

The results confirm that the bimodal specification is superior.

Table D.8: Bimodality test for ZICO

	ZICO
K=1	-2097
	(0.000)
K=2	-1947
	(1.000)

Note: This table reports the BIC values for unimodal and bimodal models, with Raftery's BIC probability in parentheses. A higher (less negative) BIC value indicates a better model fit. Raftery's BIC probability quantifies the likelihood that one model is superior to another, where a value of 1.000 suggests a strong preference for the bimodal specification.

D.4 Finite Mixture Model for optimal threshold point

Fitting an FMM with two Gaussian components, we obtain two normal distributions:

$$f_1(ZICO) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(ZICO-\mu_1)^2}{2\sigma_1^2}}$$

$$f_2(ZICO) = \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{(ZICO-\mu_2)^2}{2\sigma_2^2}}$$

where μ_1 and σ_1 are the mean and standard deviation of the zombie cluster (lower scores), and μ_2 and σ_2 are the mean and standard deviation of the non-zombie cluster (higher scores).

The intersection point $ZICO^*$ is the score at which these two distributions are equally likely, meaning:

$$f_1(ZICO^*) = f_2(ZICO^*)$$

Taking the natural logarithm of both sides and rearranging:

$$\frac{(ZICO^* - \mu_2)^2}{2\sigma_2^2} - \frac{(ZICO^* - \mu_1)^2}{2\sigma_1^2} = \log\left(\frac{\sigma_2}{\sigma_1}\right)$$

Expanding the quadratic terms:

$$\frac{ZICO^{*2} - 2\mu_2 ZICO^* + \mu_2^2}{\sigma_2^2} - \frac{ZICO^{*2} - 2\mu_1 ZICO^* + \mu_1^2}{\sigma_1^2} = 2\log\left(\frac{\sigma_2}{\sigma_1}\right)$$

Rearranging into quadratic form:

$$A \times ZICO^{*2} + B \times ZICO^* + C = 0$$

where:

$$A = \frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}$$
$$B = \frac{2\mu_2}{\sigma_2^2} - \frac{2\mu_1}{\sigma_1^2}$$
$$C = \frac{\mu_2^2}{\sigma_2^2} - \frac{\mu_1^2}{\sigma_1^2} - 2\log\left(\frac{\sigma_2}{\sigma_1}\right)$$

Since this is a standard quadratic equation, we solve for $ZICO^*$ using the quadratic formula:

$$ZICO^* = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}$$

We take the positive root, as it corresponds to the meaningful threshold within the range of zombie scores.

E Additional results and extensions

This section provides a detailed overview of manipulation strategies used to conceal misconduct. It then discusses various Bayesian updating schemes to address these tactics.

E.1 Strategic behaviors and adverse selection

We have considered various forms of manipulations, as reported in Table E.9. These behaviors have been simulated using random transformations over uniform distributions because they allow controlled, non-deterministic variation within predefined bounds while preserving the underlying structure of the original variables. The uniform distribution ensures that perturbations remain within plausible ranges without introducing artificial biases, which would be more likely with a normal distribution or other skewed transformations. Moreover, since the effects of manipulations (e.g., readability improvement, lexical diversity reduction, or citation inflation) do not follow a precise parametric form in real-world scenarios, a uniform distribution provides a flexible and interpretable method for modeling such variations without imposing unrealistic assumptions about their probability distribution. This approach also enables us to explore a broad but controlled range of possible manipulated outcomes, ensuring robustness in the analysis of strategic behaviors.

Table E.9: Avoiding detection through strategic manipulations

Manipulations	Impact	Variables impacted	Transformation
AI-generated text			
AI simplifies text but increases word repetition	1) Random noise on readability and lexical diversity	RPC1, RPC2, RPC3, LPC1, LPC2, LPC3	$X_{AI_i} = X_i + \eta_i, \quad \eta_i \sim \mathcal{N}(0, 0.3)$
	2) Reduce syntactic complexity	MML, CPS	$X_{AI_{MML}} = X_{MML} \times U(0.7, 0.9)$ $X_{AI_{CPS}} = X_{CPS} \times U(0.6, 0.8)$
	3) Increase entropy in word choice and reduce part-of-speech entropy	WENT, POS	$X_{AI_{WENT}} = X_{WENT} \times U(1.1, 1.3)$ $X_{AI_{POS}} = X_{POS} \times U(0.7, 0.9)$
	4) Reduce readability variability	RSD	$X_{AI_{RSD}} = X_{RSD} \times U(0.7, 0.9)$
GPT fluency boost			
GPT makes text more fluent but reduces lexical variety	1) Improve readability scores	RPC1, RPC2, RPC3	$X_{AI_{RPC1}} = X_{RPC1} \times U(1.2, 1.5)$ $X_{AI_{RPC2}} = X_{RPC2} \times U(1.1, 1.3)$ $X_{AI_{RPC3}} = X_{RPC3} \times U(1.05, 1.2)$
	2) Reduce lexical diversity and sentence complexity	LPC1, MML	$X_{AI_{LPC1}} = X_{LPC1} \times U(0.5, 0.7)$ $X_{AI_{MML}} = X_{MML} \times U(0.5, 0.7)$
AI hallucination (jargon overuse)			
Uses excessive jargon and fake technical sophistication but makes text unnatural	1) Increase lexical complexity by adding excessive technical jargon	LPC1, LPC2, LPC3	$X_{AI_{LPC1}} = X_{LPC1} \times U(1.3, 1.6)$ $X_{AI_{LPC2}} = X_{LPC2} \times U(1.3, 1.6)$ $X_{AI_{LPC3}} = X_{LPC3} \times U(1.2, 1.4)$
	2) Reduce readability due to overuse of complex terms	RPC1, RSD, MML	$X_{AI_{RPC1}} = X_{RPC1} \times U(0.5, 0.9)$ $X_{AI_{RSD}} = X_{RSD} \times U(1.0, 1.2)$ $X_{AI_{MML}} = X_{MML} \times U(1.1, 1.3)$
Citation stuffing			
Artificially inflating credibility by increasing reference count and self-citations	1) Increase cited references	CIT	$X_{AI_{CIT}} = X_{CIT} \times U(1.5, 1.9)$
	2) Adjust self-citation proportion	SEP	$X_{AI_{SEP1}} = X_{SEP} \times U(1.3, 1.7)$ and $X_{AI_{SEP2}} = X_{SEP} \times U(0.3, 0.6)$

Note: This table presents various manipulation strategies employed by fraudulent researchers to conceal misconduct and evade detection. It outlines the types of manipulations, their effects, the affected variables, and their transformations. The notation i in the first manipulation refers to RPC1, RPC2, RPC3, LPC1, LPC2, and LPC3, respectively. $U(.)$ denotes a uniform distribution.

As postulated, one of the most significant effects of manipulation is the downward shift in the optimal threshold for classifying a paper as a zombie. To theoretically demonstrate that manipulations influence this threshold, consider the following framework.

Manipulations systematically modify the observed features X_i of each paper, reducing the likelihood of retraction ($P(Y_i = 1 \mid X_i)$) as follows:

$$P^M(Y_i = 1) < P(Y_i = 1)$$

where $P^M(Y_i = 1)$ represents the probability of retraction after manipulation.

Since ZICO scores are inversely related to the probability of retraction, a lower $P^M(Y_i = 1)$ leads to an increase in ZICO scores:

$$ZICO_i^M > ZICO_i$$

Consequently, this results in a downward shift in the optimal classification threshold:

$$ZICO^{*M} < ZICO^*$$

The proportion of papers classified as non-zombies is given by:

$$P(ZICO_i \geq ZICO^*) = 1 - F_{ZICO}(ZICO^*)$$

where $F_{ZICO}(ZICO^*)$ is the cumulative distribution function of the ZICO score distribution. Since manipulations lower the optimal threshold $ZICO^{*M}$, the proportion of papers classified as non-zombies increases, as shown by:

$$P(ZICO_i^M \geq ZICO^{M*}) > P(ZICO_i \geq ZICO^*)$$

This implies that a greater number of papers are incorrectly classified as non-zombies, increasing the likelihood that fraudulent papers evade detection.

E.2 Bayesian editorial adaptations

E.2.1 Bayesian methods and comparison

Table E.10: Bayesian updating methods

	Formula for the corrected probability of retraction, $P(Y_i)$
Standard Bayesian	$P(Y_i = 1 \mid S_i) = \frac{P(S_i \mid Y_i)\, P(Y_i = 1)}{P(S_i)}$
Hierarchical Bayesian	$P(Y_i = 1 \mid S_i) = \frac{P(S_i \mid Y_i)\, P(\theta)}{P(S_i)}$
Time-dependent probability adjustment	$P(Y_i = 1 \mid S_i, t) = P(S_i \mid Y_i, t) \cdot \gamma_t$

Note: This table presents the Bayesian updating approaches, including standard and hierarchical methods, and a time-based probability adjustment considered in our analysis. A time-based adjustment sequence of 0.8–1.2 has been applied. The prior distribution for the retraction probability, $P(\theta)$, is modeled as a Gaussian distribution. The parameter $\theta \sim (\alpha, \beta)$ represents the underlying belief about retraction probability, where α (resp. β) corresponds to the number of past retractions (resp. non-retractions).

Table E.11: Optimal threshold adjustment via Bayesian updating

	AI-Generated	**GPT fluency boost**	**AI hallucination**	**Citation stuffing**
Standard	563.2 (54.5%)	518.0 (62.5%)	544.8 (55.1%)	565.5 (51.1%)
Hierarchical	568.7 (54.5%)	545.6 (62.5%)	559.7 (55.1%)	608.8 (34.0%)
Time-based	562.8 (52.2%)	567.6 (49.3%)	542.6 (53.4%)	578.6 (50.0%)

Note: This table presents the optimal threshold adjustments obtained through various Bayesian updating methods. In parentheses, the proportion of papers classified as safe (i.e., above the threshold) is reported.

Table E.12: AUC comparison after threshold correction

	AI-Generated	**GPT fluency boost**	**AI hallucination**	**Citation stuffing**
Bayesian	0.975	0.987	0.994	0.997
Hierarchical	0.975	0.987	0.994	0.997
Time-based	0.993	0.999	0.999	0.999

Note: This table compares AUC values after Bayesian threshold correction, assessing the impact on classification performance.

Table E.13: Wasserstein distance from the original threshold

	AI-Generated	**GPT fluency boost**	**AI hallucination**	**Citation stuffing**
Bayesian	25.86	19.92	15.42	6.420
Hierarchical	53.41	51.80	50.00	45.85
Time-based	10.81	18.15	14.96	21.91

Note: This table reports the Wasserstein distance between each corrected threshold and the original threshold, measuring the extent of divergence after Bayesian updating.

E.2.2 Effect of the time-dependent probability adjustment: proof

The dynamic factor is chosen to ensure that the proportion of zombie papers remains stable:

$$P(ZICO_i^{C,t} \geq ZICO^{*C,t}) = P(ZICO_i \geq ZICO_0)$$

where $ZICO_i^{C,t}$ is the time-based corrected ZICO, and $ZICO^{*C,t}$ the corrected ZICO threshold.

Rearranging,

$$ZICO^{*C,t} = \arg\max_{ZICO} P^{C,t}(f_1(ZICO) = f_2(ZICO))$$

Since $P^C(Y_i) > P^M(Y_i)$, it follows that:

$$ZICO_i^{C,t} < ZICO_i^M$$

which leads to an upward shift in the optimal threshold ($ZICO^{*C,t} > ZICO^{*M}$) bringing the proportion of zombies back in line with the original classification.